Printed in U.S.A.
Vol. 118, No. 4

DATA IS WHAT DATA DOES: REGULATING BASED ON HARM AND RISK INSTEAD OF SENSITIVE DATA

Daniel J. Solove

ABSTRACT—Heightened protection for sensitive data is becoming quite trendy in privacy laws around the world. Originating in European Union (EU) data protection law and included in the EU's General Data Protection Regulation, sensitive data singles out certain categories of personal data for extra protection. Commonly recognized special categories of sensitive data include racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sexual orientation and sex life, and biometric and genetic data.

Although heightened protection for sensitive data appropriately recognizes that not all situations involving personal data should be protected uniformly, the sensitive data approach is a dead end. The sensitive data categories are arbitrary and lack any coherent theory for identifying them. The borderlines of many categories are so blurry that they are useless. Moreover, it is easy to use nonsensitive data as a proxy for certain types of sensitive data.

Personal data is akin to a grand tapestry, with different types of data interwoven to a degree that makes it impossible to separate out the strands. With Big Data and powerful machine learning algorithms, most nonsensitive data give rise to inferences about sensitive data. In many privacy laws, data giving rise to inferences about sensitive data is also protected as sensitive data. Arguably, then, nearly all personal data can be sensitive, and the sensitive data categories can swallow up everything. As a result, most organizations are currently processing a vast amount of data in violation of the laws.

This Article argues that the problems with the sensitive data approach make it unworkable and counterproductive as well as expose a deeper flaw at the root of many privacy laws. These laws make a fundamental conceptual mistake—they embrace the idea that the nature of personal data is a sufficiently useful focal point for the law. But nothing meaningful for regulation can be determined solely by looking at the data itself. Data is what data does.

To be effective, privacy law must focus on harm and risk rather than on the nature of personal data. The implications of this point extend far beyond sensitive data provisions. In many elements of privacy laws, protections should be proportionate to the harm and risk involved with the data collection, use, and transfer.

AUTHOR—Eugene L. and Barbara A. Bernard Professor of Intellectual Property and Technology Law, George Washington University Law School. I would like to thank my research assistants Kimia Favagehi, Jean Hyun, Tobi Kalejaiye, and Travis Yuille for excellent work. Thanks to Ella Corren, Oscar Gandy, Bob Gellman, Vandana Gyanchandani, Heidi Liu, Paul Ohm, Paul Schwartz, Alicia Solow-Niederman, Ari Waldman, and the participants of the Privacy Law Scholars Conference (PLSC) for helpful discussions and input on this project.

INTRODUCTION

Heightened protection for sensitive data is becoming quite trendy in privacy laws around the world. These provisions in privacy laws are based on a recognition that a uniform level of privacy protection would be too simplistic. Not all situations involving personal data are equal. Some situations involve minor annoyances; others involve deleterious

consequences such as emotional distress, reputational damage, discrimination, physical threats, fraud, or the loss of a job. Some situations can even be life or death.

To avoid treating serious and minor situations uniformly, many privacy laws designate special personal data categories called "sensitive data" that receive heightened protections. With sensitive data, privacy laws offer two levels of protection: a baseline level for regular personal data and a heightened level for sensitive data. Although two levels might not be granular enough, two is certainly better than one. Commonly recognized special categories of sensitive data include racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sexual orientation and sex life, and biometric and genetic data.

Originally appearing in European Union (EU) data protection laws, sensitive data has found its way into the comprehensive privacy laws of countless countries. Long a holdout, the United States joined the bandwagon in 2020, when several state consumer privacy laws began including sensitive data. Providing heightened protections for sensitive data is sweeping across the globe. It would not be hyperbole to say that sensitive data has become one of the canonical elements of privacy laws.

This Article argues that the problems with the sensitive data approach make it unworkable and counterproductive—as well as expose a deeper flaw at the root of many privacy laws. These laws make a fundamental conceptual mistake: they embrace the idea that the nature of personal data is a sufficiently useful focal point for the law. But meaningful regulation cannot be determined solely by looking at the data itself. Data is what data does. Heightened protection of personal data should be based on the extent of harm or the risk of harm from its collection, use, or transfer.

Although it continues to rise in popularity, the sensitive data approach is a dead end. The sensitive data categories are arbitrary and lack any coherent theory for identifying them. The borderlines of many categories are so blurry that they are useless. Moreover, nonsensitive data can easily be used as a proxy for certain types of sensitive data.

Personal data is akin to a grand tapestry, with different types of data interwoven to a degree that makes it impossible to separate out the strands. The very notion that special categories of personal data can readily be demarcated fundamentally misunderstands how most personal data is interrelated and how algorithms and inferences work.

When nonsensitive data can give rise to inferences about sensitive data, many privacy laws correctly consider it to be sensitive data. Indeed, in our age of modern data analytics, it would be naïve to fail to account for inferences. The problem, however, is the rabbit hole goes all the way to

Wonderland. In the age of Big Data, powerful machine learning algorithms facilitate inferences about sensitive data from nonsensitive data. As a result, nearly all personal data can be sensitive, and thus the sensitive data categories can swallow up everything. Oddly, the laws just seem to hum along as if this problem does not exist.

The implications of this point are significant. If nearly all data is sensitive data, then most organizations are violating the EU's General Data Protection Regulation (GDPR) and many other privacy laws that have heightened protections for sensitive data.

This Article contends that privacy law requires a rethinking. To be effective, privacy law must focus on harm and risk rather than on the nature of personal data. The implications of this point extend far beyond sensitive data provisions. In many elements of privacy laws, protections should be proportionate to the harm and risk involved with the way data is collected, used, and transferred.

Currently, privacy statutes do not focus sufficiently on harm and risk. Privacy harm and risk are issues that judges and policymakers have struggled over, especially in the United States.[1] Regulating based on harm and risk is a difficult road fraught with complexity, so it is no surprise it is often the road not taken.

On the surface, the sensitive data approach appears to offer the virtue of simplicity. Even if imperfect, a simple approach might be better than a complicated one. But the sensitive data approach only appears to be simple. When examined more deeply, the sensitive data approach is quite complex because it is nearly impossible to sort data into the sensitive data categories. Enormous complexity lurks behind the mirage of simplicity.

The sensitive data approach might be defended as roughly tracking harm and risk, but the correlation is far too weak to be useful. In too many cases, sensitive data is not necessarily more harmful than nonsensitive data. The sensitive data approach has significant costs because it creates the illusion of responding to harm and risk while the most harmful and risky situations are inadequately addressed. This illusion, combined with the myth that sensitive data is simple and practical to single out, makes the sensitive data approach an elaborate hall of mirrors that leads nowhere.

This Article identifies the shortcomings of the current sensitive data approach and argues that policymakers should instead focus on harm and risk. Recognizing the practical challenges of looking at harm and risk, I argue

[1] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. REV. 793, 796–99 (2022); Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data Breach Harms*, 96 TEX. L. REV. 737, 739, 744 (2018).

that the sensitive data approach is actually more complex, impossible to implement practically, and incapable of keeping up with the exponential growth of data analytics.

This Article proceeds in several parts. Part I provides background about personal data and sensitive data. Part II examines the challenges that inferences about nonsensitive data raise for the sensitive data approach. Part III contends that focusing on the nature of the data is the wrong focus for the law. Part IV argues that the law should focus on harm and risk. Despite the complexity of this path, it is the most viable direction for privacy law to take.

I. PERSONAL DATA AND SENSITIVE DATA

Privacy laws are triggered by activities involving "personal data," which typically is defined as data involving an "identified" or "identifiable" person.[2] Once a privacy law is triggered, it typically requires a slate of protections.[3] Many privacy laws also recognize special categories of personal data called "sensitive data" that receive heightened protection. This approach is taken in recognition that not all privacy situations are the same—some are more harmful, risky, or problematic than others. Sensitive data categorization affords these situations more protections, such as restrictions on the use of data, consent requirements, and risk assessment requirements. In this Part, I discuss how personal data and sensitive data are defined, as well as how sensitive data provisions function.

A. Personal Data

All privacy laws define the type of personal data that they cover.[4] Privacy laws cannot cover all data or else they would be boundless, so they limit the scope of data they cover to data relating to people. Thus, nearly all privacy laws are triggered based on a definition of personal data.

---

[2] *See, e.g.*, Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) ch. 1, art. 4(1) [hereinafter GDPR] ("'[P]ersonal data' means any information relating to an identified or identifiable natural person . . . .").

[3] Daniel J. Solove, *The Limitations of Privacy Rights*, 98 NOTRE DAME L. REV. 975, 979–84 (2023).

[4] The term "personal data" is not uniformly used. EU law and the GDPR use the term "personal data," with many privacy laws worldwide also using this term. The privacy laws of other countries use the term "personal information." In the United States, the laws use a multitude of different terms. Some examples include: Health Insurance Portability and Accountability Act (HIPAA): protected health information (PHI); Federal Communications Act: consumer network proprietary information (CPNI); Family Educational Rights and Privacy Act (FERPA): education records; Privacy Act: personally identifiable information (PII); California Consumer Privacy Act (CCPA): personal information.

The most common definition of "personal data" is from the GDPR which defines it as "any information relating to an identified or identifiable natural person."[5] Data is *identified* if it is linked to a specific person. Data is *identifiable* if there is a chance it could be linked to a person even if it is not currently connected. The linkage can be indirect. For example, an IP address does not directly identify a person—it is just a number corresponding to a computer connected to the internet. But it is linkable to a person through internet service provider records. Even if the computer is used by many people in a household, internet activity patterns and browsing history can readily be used to determine which household member is using the computer at a particular time. An IP address thus can be identifiable to individuals, and it therefore can be considered personal data under privacy laws that include identifiability in their definitions of personal data.

Data that is not about specific people falls outside the bounds of data privacy laws. The height of Mount Kilimanjaro, the population of Brazil, or the recipe for apple pie are not personal data. Statistical data, such as the percentage of people with cancer or the number of people over the age of sixty-five, is also not personal data. If privacy laws were to regulate all data, the laws would regulate every piece of information in an encyclopedia. The laws would be overbroad to the point of uselessness.

Outside of the GDPR, how data privacy laws go about defining personal data is quite varied and complex. In the United States, several privacy laws define personal information as data that actually identifies a person.[6] For example, personal information under the U.S. Video Privacy Protection Act is defined as "information which identifies a person."[7] Data that is identifiable—that could potentially be used to identify a person—often does not count. Many data breach notification laws employ this type of definition.[8]

The problem with limiting privacy protection laws to identified data is that the identified individuals approach is obsolete in the age of Big Data. With modern data analytics, it is relatively easy to target and identify people based on data that is not directly linked to a person. For example, computer scientist Latanya Sweeney demonstrated the ability to identify 87% of people with a combination of a postal code, birth date, and gender.[9] Although the

---

[5] GDPR, *supra* note 2, art. 4(1).

[6] 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, 1828–36 (2011).

[7] Video Privacy Protection Act, 18 U.S.C. § 2710(a)(3).

[8] *See* Schwartz & Solove, *supra* note 6, at 1828–36, 1884–85.

[9] Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* 2 (Carnegie Mellon Univ. Data Priv. Working Paper, Paper No. 3, 2000), https://dataprivacylab.org/projects/identifiability/paper1.pdf [https://perma.cc/46TD-4DWY].

identified individuals approach is popular with U.S. privacy laws, hardly any other countries adopt this approach. As Graham Greenleaf notes, "almost all data privacy laws globally" define personal data "in terms of 'identifiability.'"[10] In light of its obsolescence, the identified individuals approach is starting to wane in the United States, with many of the newer laws defining personal data as relating to an identified *or an identifiable person.*[11]

Under the more common definition of personal data, which involves identified and identifiable data, the existence of the identifiability prong gives personal data a broad, open-ended, and dynamic scope. First, any data that is associated with personal data becomes personal data. For example, the fact that an unidentified person owns a dog is not personal data. But when this fact is linked to data that can identify the person, such as the person's email address, then this fact becomes personal data.

Data that can reasonably be used in combination with other data to identify a person becomes personal data too. In many circumstances, combining nonidentifiable pieces of data can identify an individual. For example, combining data that an unidentified person owns a dog with information about dog food preferences, area dog-walking services, the dogs that frequent community parks, and other local information could lead to the identification of the dog owner. Each piece of data by itself might not be enough to identify an individual, but in combination, they may. Research has shown that collecting more data increases the likelihood of identification.[12] In many cases, there is no definitive answer about whether a particular piece of data is personal data. It depends upon the availability of other pieces of data that might be combined with the particular piece of data. Moreover, it is far too simplistic to state a definitive answer as to whether certain data can be linked to a person. For many types of data, the answer depends upon the context. For example, one particular search query might not be identifiable (such as a search for a book), but another search query (such as a person's

[10] Graham Greenleaf, *California's CCPA 2.0: Does the US Finally Have a Data Privacy Act?*, 168 PRIV. L. & BUS. INT'L REP. 13 (2020).

[11] *Comparing U.S. State Data Privacy Laws vs. the EU's GDPR*, BLOOMBERG L. (July 11, 2023), https://pro.bloomberglaw.com/brief/privacy-laws-us-vs-eu-gdpr/ [https://perma.cc/K5MR-3WUZ] (comparing definitions of personal data under the GDPR and various U.S. state privacy laws and indicating that all recognize identifiable or linkable data or data that can indirectly identify a person).

[12] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1723–25 (2010) (explaining reidentification techniques that combine anonymized data sets with outside information to accurately reidentify individuals); *see also* Henry T. Greely, *The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks*, 8 ANN. REV. GENOMICS & HUM. GENETICS 343, 351–52 (2007) (stating that increasing anonymized datasets by combining multiple biobank databases makes it easier to reidentify previously anonymized information).

own name) is identifiable. The identifiability of a search query depends upon the specific query.[13]

B. Sensitive Data

Beyond protections for personal data, privacy laws around the world have heightened protection for sensitive data. Commentators refer to sensitive data as "a bedrock of modern data protection."[14] The privacy laws of many countries define certain types or categories of data, which receive greater protections than regular personal data. These types of data are referred to as "special categories of data" or "sensitive data."

Privacy laws with sensitive data provisions often have two levels of protection: one for regular personal data and one for sensitive data. A rare exception is Canada's Personal Information Protection and Electronic Documents Act (PIPEDA), which mandates protections of sensitive data that are proportionate to the sensitivity of the data without having just two levels of protection.[15] Most laws, in contrast, do not have this more granular proportional approach.

Long an outlier, the United States has lacked a recognition of sensitive data in its privacy laws. Recently, however, the United States joined the bandwagon. Since 2020, several new U.S. privacy laws—in particular, consumer privacy laws enacted by the states—started to recognize sensitive data.[16]

In this Section, I trace the emergence of sensitive data, categorize the types of data recognized as sensitive, and discuss the heightened protections and rationales afforded to it.

1. Rise of Recognition of Sensitive Data

Sensitive data initially appeared in early privacy laws from Sweden and Hesse, a German state, in the early 1970s.[17] In its influential Privacy Guidelines of 1980, the Organisation for Economic Co-operation and Development (OECD) recognized sensitive data, but merely had a barebones

---

[13] Schwartz & Solove, *supra* note 6, at 1836.

[14] Paul Quinn & Gianclaudio Malgieri, *The Difficulty of Defining Sensitive Data—the Concept of Sensitive Data in the EU Data Protection Framework*, 22 GERMAN L.J. 1583, 1587 (2021).

[15] Personal Information Protection and Electronic Documents Act, S.C. 2000, c 5, § 4.7 (Can.) ("Personal information shall be protected by security safeguards appropriate to the sensitivity of the information.").

[16] *See infra* Section I.B.1.

[17] Karen McCullagh, *Data Sensitivity: Proposals for Resolving the Conundrum*, 2 J. INT'L COM. L. & TECH. 190, 190 (2007).

account of it, without specifying how it was to be protected or what types of data should be deemed to be sensitive.[18]

In 1981, the Council of Europe's Convention No. 108 recognized sensitive data, mentioning categories including racial origins, political opinions, religious or other beliefs, health, and sexual life.[19] These categories were nonexclusive.[20]

The United Nations Guidelines for the Regulation of Computerized Data Files in 1990 recognized categories of data similar to sensitive data, yet the concept was focused narrowly on discrimination.[21] Principle 5, the "Principle of non-discrimination," provided that "data likely to give rise to unlawful or arbitrary discrimination, including information on racial or ethnic origin, colour, sex life, political opinions, religious, philosophical and other beliefs as well as membership of an association or trade union, should not be compiled."[22]

As the sensitive data approach took form, a debate arose over whether it should be an open or closed list. An open list would allow for new categories of sensitive data to be added over time; a closed list would limit the categories of sensitive data to those specified in the law, and no additional categories could be added unless the law would be amended.

In 1980, the OECD Privacy Guidelines took an open-list approach. The explanatory memo to the OECD Privacy Guidelines acknowledged that "different traditions and different attitudes by the general public have to be taken into account. Thus, in one country universal personal identifiers may be considered both harmless and useful whereas in another country they may be regarded as highly sensitive and their use restricted or even forbidden."[23] The memo further stated that "[i]t could be argued that it is both possible and desirable to enumerate types or categories of data which are per se sensitive and the collection of which should be restricted or even prohibited."[24] The memo noted that although some European legislation recognized sensitive

---

[18] Quinn & Malgieri, *supra* note 14, at 1587; *see* OECD, OECD GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA ¶ 19(a) (1980), https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/oecd_fips.pdf [https://perma.cc/5MG4-6LED] (recommending OECD Member countries "adopt appropriate domestic legislation").

[19] Explanatory Report to the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data ¶¶ 44–45, at 9, ¶ 69, at 12, Jan. 28, 1981, E.T.S. No. 108, https://rm.coe.int/16800ca434 [https://perma.cc/EAL2-QWXZ].

[20] *Id.* ¶ 48; McCullagh, *supra* note 17, at 191 ("The categories listed in Article 6 are not meant to be exhaustive. Rather, the Convention provides that a Contracting State should be free to include other categories of sensitive data.").

[21] G.A. Res. 45/95, ¶ 5 (Dec. 14, 1990).

[22] *Id.*

[23] OECD, *supra* note 18, ¶¶ 3, 45.

[24] *Id.* ¶¶ 7, 50.

data categories, "[o]n the other hand, it may be held that no data are intrinsically 'private' or 'sensitive' but may become so in view of their context and use."[25] The memo then concluded that although the Expert Group considered adopting criteria to define sensitive data, the Group ultimately "has not found it possible to define any set of data which are universally regarded as sensitive."[26]

Reflecting the view of the OECD, renowned EU data protection jurist Spiros Simitis contended in 1999 that because any personal data can be sensitive depending on the circumstances, all data should be assessed within the context of the data's use.[27] Simitis argued that sensitivity should be "no more than a mere alarm device" signaling that regulation of personal data may not be securing adequate protection.[28] According to Simitis, the primary consequence of sensitivity is "to incite a reflection process the purpose of which is to locate the shortcomings of the existing regulations and to establish the improvements needed."[29]

In contrast to the OECD Privacy Guidelines, the EU Data Protection Directive specified types of sensitive data when it mandated that all EU member nations provide heightened protections for sensitive data in 1995. The Directive required member states to prohibit "the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life."[30]

The general approach of the Directive was to set the default rule to prohibit the processing of sensitive data unless it was processed for a particular authorized reason under the law. Directives work by requiring member states to follow their particular instructions when enacting laws—each member state had to enact protections for the sensitive data categories mentioned by the Directive. The Directive did not require its list to be a closed list; it was just a minimum list and countries were allowed to include more categories. Some countries enacted laws with a closed list of the categories specified in the Directive; others enacted laws with provisions to recognize additional sensitive data categories in an open-ended way.[31]

---

[25] *Id.*

[26] *Id.* ¶¶ 7, 51.

[27] Spiros Simitis, Revisiting Sensitive Data 5 (1999) (unpublished manuscript), https://rm.coe.int/09000016806845af [https://perma.cc/BB87-NZQR].

[28] *Id.* at 8.

[29] *Id.*

[30] Council Directive 95/46, 1995 O.J. (L 281) 31, 40 (EC).

[31] McCullagh, *supra* note 17, at 197.

Following the DU Data Protection Directive, the GDPR was enacted in 2016 and added several additional categories of sensitive data to the list in the Directive, including genetic data, biometric data, and sexual orientation.[32] Unlike the Directive, which was an open list, the GDPR is a closed list. Member states cannot recognize additional categories of sensitive data.[33] However, the GDPR makes an exception and provides that "Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health."[34]

The 1995 Directive and the GDPR have had a profound influence on privacy laws beyond the EU, with many countries basing their laws on the Directive or on the laws of particular EU member countries governed by the Directive.[35] As a result, most privacy laws around the world have sensitive data protections. These laws are mixed on whether they have an open or closed list. Those with open lists typically allow the data protection authority or some other regulatory body to designate certain categories of data as sensitive on an ongoing basis.[36] Some other countries recognize any data that could lead to discrimination as sensitive.[37]

---

[32] GDPR, *supra* note 2, art. 9(1)

[33] Quinn & Malgieri, *supra* note 14, at 1589.

[34] GDPR, *supra* note 2, art. 9(4).

[35] *See* DLA PIPER, DATA PROTECTION LAWS OF THE WORLD: TURKEY 2 (2023), https://www.dlapiperdataprotection.com/system/modules/za.co.heliosdesign.dla.lotw.data_protection/functions/handbook.pdf?country-1=TR [https://perma.cc/M6AE-Q43R] (stating that the Turkish Law on the Protection of Personal Data is based primarily on the EU Directive and includes regulations limiting the processing of sensitive personal data); Law No. 25.326, Oct. 4, 2000, § 2 (Arg.) (defining sensitive data nearly identically to the EU Directive); Personal Information Protection Act, art. 23 (S. Kor.) (defining sensitive information similarly to the EU Directive); *Data Protection Laws of the World: Uruguay*, DLA PIPER (Jan. 26, 2023), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=UY [https://perma.cc/E78Y-ZJJ8] (illustrating how Uruguay's Data Protection Act defines sensitive personal data with nearly identical verbiage to the EU Directive).

[36] For example, the Personal Information Protection Law of the People's Republic of China uses open language to indicate that the list of sensitive information includes but is not limited to the examples presented in the law. The law provides: "'Sensitive personal information' is personal information that once leaked or illegally used, may easily lead to the infringement of the personal dignity of a natural person or may endanger his personal safety or property . . . ." Zhonghua Renmin Gongheguo Geren Xinxi Baohu Fa (中华人民共和国个人信息保护) [Personal Information Protection Law of the People's Republic of China] (promulgated by the Standing Comm. Nat'l People's Cong., Aug. 20, 2021), art. 28 § 2, 2021 Standing Comm. Nat'l People's Cong. Gaz. 13 (China). The law then lists some examples, but the list does not appear to be exclusive.

[37] *See Data Protection Laws of the World: Japan*, DLA PIPER (Jan. 1, 2023), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=JP [https://perma.cc/X28N-C9H8] (defining sensitive data, in part, to be any information "that might cause the person to be discriminated against"); *see also Data Protection Laws of the World: Colombia*, DLA PIPER (Jan. 28, 2023), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=CO [https://perma.cc/8RP4-

The United States was long a holdout on recognizing sensitive data. But starting in 2020, state consumer privacy laws began including heightened protections for sensitive data. The first of these consumer privacy laws was the California Consumer Privacy Act (CCPA) of 2018. Originally, the CCPA did not recognize sensitive data, but heightened protection for sensitive data was added by a referendum in 2020 called the California Privacy Rights Act (CPRA).[38] Subsequently, several states passed consumer privacy laws inspired by California and included heightened sensitive data protections, including Colorado, Connecticut, Florida, Indiana, Iowa, Montana, Tennessee, Texas, Utah, and Virginia.[39] While U.S. law now provides for sensitive data protections, the categories of sensitive data vary across privacy laws, with all taking a closed-list approach so far.

2. Types of Data Recognized as Sensitive

Most laws that have heightened protections for sensitive data define it in terms of specific categories. For example, under the GDPR, sensitive data includes the following special categories of personal data:

- racial or ethnic origin;
- political opinions;
- religious or philosophical beliefs;
- trade-union memberships;
- health;
- sex life or sexual orientation;
- genetic data; and
- biometric data.[40]

In a 2019 analysis of sensitive data definitions from 112 countries, the most commonly recognized categories of sensitive data include the types

K22C] (defining sensitive data in Columbia as "any data that affects its owner's intimacy or whose improper use might cause discrimination"); *Data Protection Laws of the World: Ecuador*, DLA PIPER (Jan. 26, 2023), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=EC [https://perma.cc/99JQ-BQ4F] (including data "whose improper processing may give rise to discrimination" in the definition of sensitive data); Ley No. 787, 21 Mar. 2012, Ley de Protección de Datos Personales [Law on Personal Data Protection] ch. 1, art. 3(g), LA GACETA, DIARIO OFICIAL [L.G.], 29 Mar. 2012 (Nicar.) (defining sensitive data, in part, as information that may be a reason for discrimination).

[38] California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.135 (West 2023).

[39] For examples of some of these new state laws, see COLO. REV. STAT. § 6-1-1303(24) (2021); 2023 Conn. Pub. Acts No. 22-15 § 1(27); UTAH CODE ANN. § 13-61-101(32) (West 2023); VA. CODE ANN. § 59.1-575 (2023). *See also* Nancy Libin, Michael T. Borgia, John D. Seiver, David L. Rice & Patrick J. Austin, *Florida Digital Bill of Rights Signed into Law*, DAVIS WRIGHT TREMAINE LLP (June 8, 2023), https://www.dwt.com/blogs/privacy--security-law-blog/2023/06/florida-digital-bill-of-rights-data-privacy [https://perma.cc/BAN9-3MNA].

[40] GDPR, *supra* note 2, art. 9.

defined by the GDPR.[41] A major divergence is that the laws of many other countries include criminal records as sensitive data whereas the GDPR does not (although the GDPR provides special protection for criminal records).[42] Other commonly recognized types of sensitive data include credit information and identification numbers or documents.[43] Some countries define sensitive data as including data related to private life, personal habits, or private relationships.[44]

A few types of data that are occasionally recognized as sensitive include abnormal addiction, age, child adoption, contact information, home address, domestic violence information, education, gender, geolocational, and social status.[45] Turkey uniquely recognizes clothing as sensitive data, likely because of the possibility that clothing can give rise to inferences about religion.[46]

In the United States, most state consumer privacy laws passed thus far recognize the following categories of sensitive data:

- racial or ethnic origin;
- religious beliefs;
- mental or physical health diagnosis;
- sexual orientation; and
- genetic or biometric data.

Other categories of sensitive data that are commonly recognized (but not as widely as the list above) include citizenship or immigration status and personal data collected from a known child.

The CCPA also recognizes the following types of data as sensitive:

- Social Security, driver's license, state identification card, or passport number;
- account log-in details;
- financial account, debit card, or credit card number;
- philosophical beliefs;
- union membership; and

---

[41] K Royal, Sensitive Data Chart (Sept. 13, 2019) (on file with *Northwestern University Law Review*).

[42] *Id.* GDPR Article 10 permits the processing of personal data relating to criminal convictions and offenses when the law of a member state authorizes the processing and provides "for appropriate safeguards for the rights and freedoms of data subjects." GDPR, *supra* note 2, art. 10.

[43] K Royal, *supra* note 41.

[44] *Id.*

[45] *Id.*

[46] *See Data Protection Laws of the World: Turkey*, DLA PIPER (Jan. 12, 2023), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=TR [https://perma.cc/F8Y2-Z8C7] (showing that the Turkish LPPD includes clothing in its definition of sensitive personal data).

- contents of mail, email, and text messages, unless the business is the intended recipient of the communication.[47]

Unlike the GDPR, the U.S. state privacy laws do not recognize political opinions as sensitive data. Additionally, most U.S. state laws (except for the CCPA) fail to recognize philosophical beliefs as sensitive data, contrary to the GDPR.

Overall, privacy laws have significant overlap in the categories of data they recognize as sensitive, but they also have many differences. The result is a rather complicated landscape from jurisdiction to jurisdiction. To comply with the law, organizations must classify their personal data (a practice known as "data mapping"), identifying which data is sensitive because it must be treated differently. With more than 70% of the 194 countries around the world having comprehensive privacy laws (most of which include sensitive data),[48] plus laws in different U.S. states, mapping which data is sensitive is a complicated task. Even categories that laws recognize in common may have slight differences, such as "health" data under the GDPR versus "mental or physical health diagnosis" under the Virginia Consumer Data Protection Act (VCDPA).[49] Accordingly, complying with these laws is quite challenging.

3. Types of Heightened Protections for Sensitive Data

Sensitive data receives heightened protections under the laws that recognize it. These protections typically involve restrictions on processing the data, more frequent requirements to have express consent to process the data, and a requirement to carry out a privacy risk assessment before processing.

Sensitive data is primarily protected by requiring express consent to process data under certain circumstances. Under the GDPR, consent must be a "freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him

---

[47] CAL. CIV. CODE § 1798.135 (West 2023).

[48] According to the United Nations Conference on Trade and Development, 137 of 194 countries have comprehensive data privacy laws. *Data Protection and Privacy Legislation Worldwide*, UNITED NATIONS CONF. ON TRADE & DEV. (Dec. 14, 2021), https://unctad.org/page/data-protection-and-privacy-legislation-worldwide [https://perma.cc/M8C7-PKHB]. According to Graham Greenleaf, a professor who tracks global privacy laws, there are 157 countries with comprehensive privacy laws as of mid-March 2022. Graham Greenleaf, *Now 157 Countries: Twelve Data Privacy Laws in 2021/22*, 176 PRIV. L. & BUS. INT'L REP., Apr. 2022, at 1.

[49] *Compare* GDPR, *supra* note 2, art. 9 (stating "data concerning health"), *with* VA. CODE ANN. § 59.1-575 (2023) (stating "mental or physical health diagnosis").

or her."[50] Consent is not the only way to process sensitive data, but the structure of the GDPR (and other laws modeled on the GDPR or its predecessor, the EU Data Protection Directive) leads to consent playing a much greater role in the processing of sensitive data. How the GDPR achieves this is quite complicated, and it requires some background to understand.

Under the GDPR, personal data cannot be collected or processed without a "lawful" basis—a permissible reason specified in the law.[51] The GDPR specifies six lawful bases:

> (1) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
> (2) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
> (3) processing is necessary for compliance with a legal obligation to which the controller is subject;
> (4) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
> (5) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
> (6) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.[52]

Personal data can be collected and processed only for one of these bases: consent, contract, legal compliance, vital interests, public interest, and legitimate interests. In practice, the main basis used to process personal data without consent is for legitimate interests.

Sensitive data requires an additional step—another legal basis—to process. Article 9 of the GDPR, which governs sensitive data, begins with a general prohibition on processing sensitive data:

> Processing of 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

[50] GDPR, *supra* note 2, art. 4(11).
[51] *Id.* art. 6.
[52] *Id.*

> identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.[53]

Then, Article 9 lists ten exceptions as allowable legal bases to process sensitive data.[54] Because these exceptions are long and wordy, I summarize them below as follows:

- consent of the data subject;
- employment and social security purposes;
- protect people's vital interests;
- charitable activities;
- the data subject made the data publicly available;
- legal claims;
- public interest;
- healthcare uses;
- public health purposes; and
- research or statistical purposes.

The GDPR essentially sets up a series of two hurdles in order to process sensitive data. The first hurdle involves a legal basis to process personal data, and the second hurdle involves a legal basis to process sensitive data.

The legal bases to process sensitive data overlap with some of the six legal bases to process regular personal data, such as consent, legal claims, public interest, and vital interests.[55] In most cases, once the first hurdle is cleared, the second hurdle can also be cleared. If data is processed with consent, then it will clear the first and second hurdles since consent is a legal basis to process personal data and sensitive data. Some of the bases to process sensitive data are additional ones that are not on the list of six to process regular personal data.

There is one very notable omission from the bases to process sensitive data: legitimate interests. Thus, in practice, the main difference between personal data and sensitive data is that the legitimate interests basis cannot be used for sensitive data. The inability to process sensitive data for legitimate interests is a significant limitation because legitimate interests is a frequently used justification to process data without consent. The other legal bases to process data without consent are fairly narrow, making legitimate interests a preferred basis to employ. Many organizations want to use personal data to market, monetize, influence, or persuade, among other things, and these reasons do not fit with the other legal bases. Beyond

---

[53] *Id.* art. 9(1).

[54] *Id.* art. 9(2).

[55] *Id.* art. 6.

obtaining express consent, which can be quite difficult, the main way to process is through the legitimate interests legal basis. Thus, without the legitimate interests legal basis to process sensitive data, organizations often must resort to obtaining consent before processing.

Other special protections for sensitive data in the GDPR include the requirement to appoint a data protection officer (DPO) and to conduct a data protection impact assessment (DPIA) when processing a "large scale of special categories of data."[56]

In a few ways, the GDPR has lessened the difference between the protections for regular personal data and sensitive data. Prior to the GDPR, the EU Data Protection Directive's consent requirement for processing sensitive data was more stringent than the consent required to process regular personal data.[57] But the GDPR essentially removed any meaningful difference between the nature of consent required for personal versus sensitive data.[58] Additionally, the GDPR does not allow member states to add further protections to sensitive data except for genetic, biometric, and health data.[59]

Moving beyond the GDPR, other countries provide similar protections for sensitive data. In the United States, the state consumer privacy laws provide heightened protections for sensitive data. For example, the CCPA states that sensitive data "shall be treated as personal information for purposes of all . . . sections of [the CCPA]" except when it is gathered or processed for "the purpose of inferring characteristics about a consumer."[60] The CCPA only provides a limited protection of sensitive data, allowing consumers to limit the use and disclosure of sensitive data to what is "necessary to perform the services or provide the goods reasonably expected by an average consumer who requests those goods or services."[61] Essentially, the CCPA provides an opt out, as sensitive data may be processed unless an individual objects. Unlike the CCPA, the VCDPA and the Colorado Privacy Act (CPA) require express consent and a data protection impact assessment to process sensitive data.[62]

---

[56] *Id.* art. 6.1(c), 35.3(b).

[57] Council Directive 95/46, art. 8, 1995, O.J. (L 281) (EC).

[58] Quinn & Malgieri, *supra* note 14, at 1601–02.

[59] *Id.* at 1589; GDPR, *supra* note 2, art 9(4).

[60] CAL. CIV. CODE § 1798.121(d) (West 2023).

[61] *Id.* § 1798.121(a).

[62] VA. CODE ANN. §§ 59.1-578(A)(5), 580(A)(4) (2023); COLO. REV. STAT. §§ 6-1-1308(d)(7), 1309 (2021) (requiring controllers to obtain consent and conduct data protection assessments before processing sensitive information).

4. Rationale for Heightened Protection of Sensitive Data

Many laws offer no particular rationales for why they protect sensitive data. But at the most basic level, sensitive data is rooted in a recognition that not all situations involving personal data are the same. The sensitive data approach focuses on the type of personal data involved to distinguish situations that should be afforded heightened protection.

Under this view, some personal data can seem quite innocuous while other data can be very revealing, embarrassing, or damaging to one's reputation. For example, on the surface level, data that a person is wearing a blue shirt does not appear to be particularly harmful or revealing. But other personal data, such as the fact that a person has a fatal disease, would be harmful or revealing. Some diseases carry stigma, so the person could be embarrassed or suffer reputational harm if this data is disclosed. The person could also suffer discrimination, finding it hard to be hired for a job or to receive a loan.

EU academics Paul Quinn and Gianclaudio Malgieri observe that EU law sometimes uses instrumental rationales for protecting sensitive data but other times views protecting sensitive data as an end in and of itself.[63] Two instrumental rationales predominate: (1) to protect against risks to fundamental rights and freedoms and (2) to protect against unlawful discrimination. As the GDPR provides at Recital 51: "Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms."[64] The European Court of Justice has quoted this language when explaining why sensitive data is protected more stringently.[65] Likewise, the Council of Europe offered a justification for sensitive data in its explanatory report on the Modernized Convention 108 on Automatic Processing of Personal Data. Sensitive data can involve "a potential risk of discrimination or injury to an individual's dignity or physical integrity, where the data subject's most intimate sphere, such as his or her sex life or sexual orientation, is being affected, or where processing of data could affect the presumption of innocence."[66] Finally, the United Nations Guidelines for the Regulation of

[63] Quinn & Malgieri, *supra* note 14, at 1585–87.

[64] GDPR, *supra* note 2, recital 51.

[65] Case 184/20, OT v. Vyriausioji tarnybinės etikos komisija, 2022 E.C.R ¶ 51.

[66] Explanatory Report to the Protocol Amending the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data art. 6, ¶ 55, Oct. 2018, E.T.S. No. 223, https://rm.coe.int/cets-223-explanatory-report-to-the-protocol-amending-the-convention-fo/16808ac91a [https://perma.cc/G5SZ-J77A].

Computerized Data Files in 1990 included heightened protection of sensitive data because it created a risk of "unlawful or arbitrary discrimination."[67]

* * *

Sensitive data is now fully established as a key element of many privacy laws. Even the United States, long a holdout, has now joined the chorus. Unfortunately, despite the growing popularity of sensitive data, it is wrongheaded. In the remainder of this Article, I argue that the sensitive data approach is flawed and doomed—it cannot be fixed. Not only is sensitive data unworkable, but it is also undesirable. It is based on a fundamental error that reverberates throughout many privacy laws.

II. THE POWER OF INFERENCE: NEARLY ALL DATA IS SENSITIVE DATA

Today, Big Data employs a legion of sophisticated algorithms to analyze data, many of which involve machine learning, where they evolve as they are fed increasing quantities of data.[68] Inferences about sensitive data can readily be made from nonsensitive data.[69] Race can be inferred from where a person lives. Religion can be inferred from location or eating patterns. Philosophical beliefs can be inferred from reading habits. Political beliefs can be inferred from nearly anything, as an increasing array of issues and behaviors are politicized. We live in what law professor Alicia Solow-Niederman aptly calls an "inference economy."[70]

Under several major privacy laws, inferences count as sensitive data. Any personal data from which sensitive data can be inferred will also be deemed to be sensitive data. The problem, though, is that the implications are far greater than currently recognized. This Part explores how in an age of inference, nearly all regular personal data can, either in isolation or combination, give rise to inferences about sensitive data. Research continually and emphatically demonstrates how readily inferences about sensitive data can be made. As algorithms grow more sophisticated and

---

[67] Guidelines for the Regulation of Computerized Personal Data Files, G.A. Res. 45/95, ¶ 5 (Dec. 14, 1990).

[68] *See* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 75–77 (2016).

[69] As Yuki Matsumi aptly argues, there are different types of inferences routinely made about people. Some involve inferences about the state of things in the present. Others involve predictions about the future, which are not verifiable. Hideyuki Matsumi, *Predictions and Privacy: Should There Be Rules About Using Personal Data to Forecast the Future?*, 48 CUMB. L. REV. 149, 150 (2017).

[70] Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 NW. U. L. REV. 357, 361 (2022) (defining an "inference economy" as one "in which organizations use available data collected from individuals to generate further information about both those individuals and other people").

consume vaster quantities of data, they will be able to make even more inferences about sensitive data—inferences that are quite unexpected and difficult to anticipate. Privacy laws often gloss over this problem, but this is not a minor glitch to be tweaked. It is a difficulty that makes the sensitive data approach tremendously complicated as well as essentially unworkable.

A. Inferences Count

As discussed in Section I.A, identifiable data is considered to be personal data under the definitions of many privacy laws. Essentially, nonidentified data is identifiable when inferences can be made about it that link it to a person. The same principle applies to sensitive data. Under the GDPR and the laws of other countries, data that could give rise to inferences about sensitive data is deemed to be sensitive data too.[71] The European Data Protection Board (EDPB) has stated that "[p]rofiling can create special category data by inference from data which is not special category data in its own right but becomes so when combined with other data."[72]

According to EU guidance, health data includes data that "can be used in itself or in combination with other data to draw a conclusion about the actual health status or health risk of a person."[73] Thus, health data "also includes data about the purchase of medical products, devices and services, when health status can be *inferred* from the data."[74] In another guideline, the Article 29 Working Party, the predecessor to the EDPB, noted that "it may be possible to infer someone's state of health from the records of their food shopping combined with data on the quality and energy content of foods."[75]

In a case from 2022, the European Court of Justice (CJEU) held that data giving rise to inferences about sensitive data is also sensitive data under the GDPR.[76] The case involved a Lithuanian law that required people receiving public funds to submit a declaration of interest, which included

---

[71] Article 29 Data Protection Working Party, *Advice Paper on Special Categories of Data ("Sensitive Data")*, at art. 8(1) (2011), https://ec.europa.eu/justice/article-29/documentation/other-document/files/2011/2011_04_20_letter_artwp_mme_le_bail_directive_9546ec_annex1_en.pdf [https://perma.cc/R53R-NXVK] (noting that sensitive data includes "not only data which by its nature contains sensitive information . . . but also data from which sensitive information with regard to an individual can be concluded").

[72] Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, at art. 9, 17/EN WP251rev.01 (2018), https://ec.europa.eu/newsroom/article29/items/612053/en [https://perma.cc/YW6D-87ED].

[73] *Id.*; Annex—Health Data in Apps and Devices, at 1, 2, Article 29 Data Protection Working Party, (2015), https://ec.europa.eu/justice/article-29/documentation/other-document/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf [https://perma.cc/FZ4M-ULRE].

[74] Annex—Health Data in Apps and Devices, *supra* note 73, at 1, 2 (emphasis added).

[75] Article 29 Data Protection Working Party, *supra* note 72, at 15.

[76] Case 184/20, OT v. Vyriausioji tarnybinės etikos komisija, 2022 E.C.R ¶ 117.

information about their spouses, partners, or cohabitants. These declarations were published online. The plaintiff challenged the requirement as a violation of the GDPR because it could lead to inferences about the plaintiff's sexual orientation—one of the types of sensitive data under the GDPR.

The CJEU held:

> [T]he publication, on the website of the public authority responsible for collecting and checking the content of declarations of private interests, of personal data that are liable to disclose indirectly the political opinions, trade union membership or sexual orientation of a natural person constitutes processing of special categories of personal data, for the purpose of those provisions.[77]

The court noted that "it is possible to deduce from the name-specific data relating to the spouse, cohabitee or partner of the declarant certain information concerning the sex life or sexual orientation of the declarant and his or her spouse, cohabitee or partner."[78] Accordingly, publishing data "liable to disclose indirectly the sexual orientation of a natural person constitutes processing of special categories of personal data, for the purpose of those provisions."[79] The court reasoned that the effectiveness of heightened protection for sensitive data would be undermined if data giving rise to inferences about sensitive data were not included.[80]

One issue is whether the sensitivity of inference-producing data should be viewed objectively based on the possibility of making inferences or subjectively based on the stated intentions of the data controller. Paul Quinn and Gianclaudio Malgieri note that EU law is inconsistent on this question and further note difficulties in both approaches.[81] Professors Sandra Wachter and Brent Mittelstadt note that the Article 29 Working Party has not addressed the question directly. But it has recognized some types of data can be objectively designated as sensitive, without regard to the intentions of those seeking to process data.[82]

---

[77] *Id.*

[78] *Id.* ¶ 119.

[79] *Id.* ¶ 128.

[80] *Id.* ¶ 127.

[81] Quinn & Malgieri, *supra* note 14, at 1591–96 (using the terms "contextual approach" to refer to the objective method and "purposeful approach" to refer to the subjective method).

[82] Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494, 565–66.

In the United States, the CCPA does not directly refer to sensitive data but recognizes inferences as a form of personal data.[83] The CCPA's broad definition of "personal information" includes "[i]nferences drawn from any of the information identified in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes."[84] This definition is one of the first to explicitly encompass inferences.[85] An opinion by the California Office of the Attorney General explains:

> Inferences are one of the key mechanisms by which information becomes valuable to businesses, making it possible to target advertising and solicitations, and to find markets for goods and services. In some cases, marketing tactics are so tailored that they feel intrusive or unsettling to consumers. In other cases, consumers may never know that they are being excluded from seeing certain ads, offers, or listings based on discriminatory automated decisions. In almost every case, the source as well as the substance of these inferences is invisible to consumers. In light of all these circumstances, inferences appear to be at the heart of the problems that the CCPA seeks to address.[86]

It would be odd for California to recognize inferences about personal data but not inferences about sensitive data, as everything stated in the opinion about personal data inferences would be relevant to sensitive data inferences.

The draft rules to the CPA are even more explicit than the CCPA, as they define "sensitive data inferences" to include inferences about sensitive data made "alone or in combination with other data."[87]

Categorizing data that gives rise to inferences about sensitive data as sensitive is the only coherent approach privacy laws can take. Otherwise, sensitive data protections would be meaningless because inferences from nonsensitive data could readily be used, thus allowing relatively easy navigation around any restrictions for sensitive data.

---

[83] Jordan M. Blanke, *Protection for 'Inferences Drawn': A Comparison Between the General Data Protection Regulation and the California Consumer Privacy Act*, 1 GLOB. PRIV. L. REV. 81, 89–90 (2020) [hereinafter Blanke, *Comparison*]; Jordan Blanke, *The CCPA, "Inferences Drawn," and Federal Preemption*, 29 RICH. J.L. & TECH. 53, 61 (2022).

[84] CCPA, CAL. CIV. CODE § 1798.140(v)(1)(K) (West 2023).

[85] Blanke, *Comparison*, *supra* note 83, at 81.

[86] Cal. Off. of the Att'y Gen., No. 20-303, Opinion Letter on Inferences Under the CCPA 13 (Mar. 10, 2022).

[87] Colorado Privacy Act Rules, COLO. CODE REGS. § 904-3-2.02 (2023).

B. Inference-A-Rama

In today's world of sophisticated data analytics, it is quite easy to infer sensitive data from nonsensitive data.[88] In the Section above, I argued that if privacy laws failed to recognize data as sensitive when it could give rise to inferences about sensitive data, this would make a mockery of sensitive data protections. On the flip side, however, if sensitive data includes data giving rise to inferences about sensitive data, then sensitive data would swallow up nearly all personal data.

A few relatively obvious examples of ways to infer sensitive data from nonsensitive data include:

- Data about patterns of electricity use can be used to infer that a person is an Orthodox Jew because Orthodox Jews do not use electricity on Saturdays.
- Data about food consumption can be used to infer religion, as some Muslims, Jews, Hindus, and members of other faiths do not eat particular foods.
- Data about food consumption can be used to infer health conditions, as particular diets are tailored to particular conditions. For example, celiac disease and diabetes are linked to gluten-free and sugar-free diets.
- Location data can be used to determine the religious or political institutions a person visits.

An extensive body of research shows how easily and accurately algorithms can make inferences about sensitive data from nonsensitive data. An examination of 327 studies about inferences found that inferable attributes include gender, age, politics, location, occupation, race and ethnicity, family and relationships, education, income, health, religion, sexual orientation, and social class.[89]

It is worth taking time to explore some examples of the inferences that are possible. Below, I examine how readily inferences can be made about many common types of sensitive data, including health, political beliefs, sexuality, and race and ethnicity.

---

[88] Wachter & Mittelstadt, *supra* note 82, at 561, 564 (noting the distinction between sensitive and nonsensitive data "is increasingly strained in the era of Big Data analytics"); Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1013 (2017) ("Big Data potentially undermines the entire distinction between these categories."); Quinn & Malgieri, *supra* note 14, at 1590–91 ("Taken with never ending increases in computing power and the increasing ease of sharing and combining disparate datasets, more and more data is arguably becoming of a sensitive nature."); Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1170 (2015) ("Many big data techniques focus on drawing accurate inferences about people from data. As these techniques increase, we might expand the use and breadth of categories of inferentially sensitive information.").

[89] Joanne Hinds & Adam N. Joinson, *What Demographic Attributes Do Our Digital Footprints Reveal? A Systematic Review*, 13 PLOS ONE, Nov. 28, 2018, at 1, 5.

1. Health

In the EU, any medical data that can give rise to an inference "about the actual health status or health risk of a person" constitutes health data.[90] If this is true, then nearly everything constitutes health data. Health data can readily be inferred from countless other types of nonsensitive data. Nearly everything people do, buy, and eat can affect health, as can gender, age, race, ethnicity, and location.

Several studies show how inferences about health can be made based on social media data. An analysis of Facebook likes inferred drug use with 65% accuracy.[91] Based on users' posts on the social media site Reddit, other researchers developed a model which purported to identify mental illnesses such as depression, bipolar disorder, schizophrenia, and borderline personality disorder.[92] Another group of researchers developed a model to predict the likelihood of postpartum depression based on Facebook data prior to giving birth. The model focused on reduction in "social activity and interaction on Facebook."[93] Information as mundane as the frequency of activity on a social media site can be the critical piece of data that lights up an algorithm.

Health data can be inferred from buying habits. One of the most famous incidents, chronicled by Charles Duhigg in the *New York Times Magazine* in 2012, involved an algorithm created by the store Target to identify women who were pregnant based on their buying habits. The algorithm was designed to detect pregnancy before women started to buy baby products in order to advertise to them early on.[94] When the father of a teenage girl saw many ads from Target for baby products, he complained to the store that the ads were being sent to the wrong house. But he later found out that his daughter was pregnant.[95]

---

[90] Article 29 Data Protection Working Party, *supra* note 74.

[91] Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PNAS 5802, 5802–03 (2013).

[92] Jina Kim, Jieon Lee, Eunil Park & Jinyoung Han, *A Deep Learning Model for Detecting Mental Illness from User Content on Social Media*, 10 SCI. REPS. 11846 (2020). It should be noted that individual verification was not provided for whether participants actually suffered from the mental illnesses represented in their posts.

[93] Munmun De Choudhury, Scott Counts, Eric J. Horvitz & Aaron Hoff, *Characterizing and Predicting Postpartum Depression from Shared Facebook Data*, *in* ASS'N FOR COMPUTING MACH., CSCW '14: PROCEEDINGS OF THE 17TH ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK & SOCIAL COMPUTING 625, 626 (2014).

[94] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES MAG. (Feb. 16, 2012), https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [https://perma.cc/VZ5L-QSTV].

[95] *Id.*

Duhigg provided more detail about the story in his book, *The Power of Habit*, which was published in the same year as his article.[96] A key piece of data used by the algorithm was that pregnant women "were buying unusually large quantities of unscented lotion around the beginning of their second trimester."[97] Additionally, they bought vitamins, scent-free soap, and cotton balls.[98]

When the story broke, it became the prime example of the privacy problems of data analytics.[99] It has been cited countless times. Unfortunately, the privacy lessons from the case were lost on the creators of the algorithm. One Target executive explained to Duhigg what they learned:

> With the pregnancy products, though, we learned that some women react badly. Then we started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random. . . . And we found out that as long as a pregnant woman thinks she hasn't been spied on, she'll use the coupons. She just assumes that everyone else on her block got the same mailer for diapers and cribs. As long as we don't spook her, it works.[100]

As most privacy experts know, this is the wrong lesson. The right lesson is: *Don't use data analytics in invasive ways to find out facts people don't reveal*. But the Target executives twisted this lesson to another: *Conceal your invasive data analytics so that people aren't aware of what you're doing*.

Ultimately, there is another lesson in this story: nonsensitive data, such as mundane purchases for lotion, soap, and cotton balls, can be used to infer sensitive data about health. With sophisticated data analytics, even rather innocuous information can reliably be used to infer sensitive data.

2. Political Opinions

In today's highly politicized environment, even innocuous products have a political valence. Mike Lindell, the head of MyPillow, prominently promoted the lie that the 2020 election was stolen and that Donald Trump

---

[96] CHARLES DUHIGG, THE POWER OF HABIT: WHY WE DO WHAT WE DO IN LIFE AND BUSINESS 182–97 (2012).

[97] *Id.* at 194.

[98] *Id.*

[99] *See* Blanke, *supra* note 83, at 82 (characterizing the Target incident as "probably the most widely publicized episode illustrating both the effect and the accuracy of predictive analysis"); Damian Fernandez-Lamela, *Lessons from Target's Pregnancy Prediction PR Fiasco*, LINKEDIN (June 16, 2014), https://www.linkedin.com/pulse/20140616204813-2554671-lessons-from-target-s-pregnancy-prediction-pr-fiasco/ [https://perma.cc/MGW3-ZLAX] ("A media and public relations storm followed, as many people were outraged at the idea of a company figuring out a highly personal situation like being pregnant.").

[100] DUHIGG, *supra* note 96, at 209–10.

was the winner.[101] This led to calls to boycott Lindell's pillows. Lindell took a "commonplace object and imbued it with a political ideology."[102] After Lindell's public embrace of Trump and the election lies, buying a pillow from MyPillow has a new meaning and political valence. What was once an innocuous purchase is now something that can be used to infer political opinions. Of course, not all purchasers of MyPillow pillows hold the same beliefs as Lindell, but correlations can grow stronger as the meaning of certain actions change over time based on circumstances. This example provides two key lessons: (1) inferences about political opinions can be made from seemingly innocuous data such as pillow purchases and (2) the landscape is constantly changing, as different products and actions take on different political significance.

Social media activity provides ample data from which inferences about political opinions can be derived. In a study involving the Facebook likes of nearly sixty thousand people, researchers could infer political party affiliation 85% of the time.[103] In one study from the United Kingdom, researchers developed an algorithm that could correctly identify political leanings 86% of the time from Twitter activity.[104] In another study, an analysis of people's Twitter activity, such as retweets and use of hashtags, among other things, enabled political affiliation to be correctly identified 91% of the time.[105]

In one of the most notorious examples of making inferences about political opinions, Cambridge Analytica mined Facebook users' data to profile them and target political advertisements towards them to vote for Donald Trump.[106] Cambridge Analytica enticed people to take a personality quiz.[107] Cambridge Analytica then gained access to the personal data of

---

[101] *See* Elizabeth Chang, *MyPillow Boycott: How a Product Can Spark an Identity Crisis*, WASH. POST (Feb. 12, 2021, 10:38 AM), https://www.washingtonpost.com/lifestyle/wellness/my-pillow-lindell-boycott-customers/2021/02/12/7399aaa4-6af1-11eb-9ead-673168d5b874_story.html [https://perma.cc/HXS2-2T53].

[102] *Id.*

[103] Kosinski et al., *supra* note 91, at 5802–03.

[104] Antoine Boutet, Hyoungshick Kim & Eiko Yoneki, *What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election*, 6 PROC. INT'L AAAI CONF. ON WEBLOGS & SOC. MEDIA 411 (2012).

[105] Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini & Filippo Menczer, *Predicting the Political Alignment of Twitter Users*, 2011 IEEE INT'L CONF. ON PRIV., SEC., RISK & TR. & IEEE INT'L CONF. ON SOC. COMPUTING 192.

[106] *See* Matthew Rosenberg, Nicholas Confessore & Carole Cadwalladr, *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. TIMES (Mar. 17, 2018), https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html [https://perma.cc/RFP8-YA43].

[107] *Id.*

individuals from their friends who took the quiz—potentially as many as eighty-seven million individuals.[108]

In one study, researchers were able to show that "information about the user's activity in nonpolitical discussion forums alone, can very accurately predict political ideology."[109] For example, the frequent use of the word "feel" was correlated with "economically left wing views."[110]

One study went beyond binary political characterizations in the United States to develop a seven-point spectrum to classify people and include people who were moderate or neutral. Using Twitter data, the researchers were able to obtain a more granular portrait of people's political leanings.[111]

3. Sexual Orientation

Researchers have also attempted to infer sexual orientation from social media activity. In one study, researchers claimed to infer sexuality 88% of the time by analyzing male Facebook likes.[112] One's social network of friendships can also be used to infer sexual orientation.[113]

In another study, researchers used an algorithm to identify sexual orientation based on facial images. The researchers claimed to correctly identify the sexual orientation for 81% of men and 71% of women. Humans looking at the same images were much less accurate, only guessing 61% correctly for men and 54% for women. According to the researchers, the algorithm's accuracy increased with more photos: with five photos available for a given participant, the algorithm was correct for 91% of men and 83% of women.[114]

---

[108] Daniel Susser, Beate Roessler & Helen Nissenbaum, *Online Manipulation: Hidden Influences in a Digital World*, 4 GEO. L. TECH. REV. 1, 10 (2019).

[109] Michael Kitchener, Nandini Anantharama, Simon Angus & Paul A. Raschky, Predicting Political Ideology from Digital Footprints 3 (May 2022) (unpublished manuscript), https://doi.org/10.48550/arXiv.2206.00397 [https://perma.cc/A822-JQEM].

[110] *Id.*

[111] Daniel Preoțiuc-Pietro, Ye Liu, Daniel J. Hopkins & Lyle Ungar, *Beyond Binary Labels: Political Ideology Prediction of Twitter Users*, 55 PROCS. ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 729 (2017).

[112] Kosinski et al., *supra* note 91, at 5802–03; *see also* Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov & David Stillwell, *Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines*, 70 AM. PSYCH. 543, 547–49 (2015) (discussing the ability and drawbacks to using Facebook likes for psychological research).

[113] Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook Friendships Expose Sexual Orientation*, FIRST MONDAY (Oct. 5 2009), https://firstmonday.org/ojs/index.php/fm/article/view/2611/2302 [https://perma.cc/MX3F-2AKV].

[114] *See* Yilun Wang & Michal Kosinski, *Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images*, 114 J. PERSONALITY & SOC. PSYCH. 246 (2018).

This study sparked significant criticism about its ethics as well as its methods and accuracy.[115] Researchers and companies are increasingly making algorithmic inferences despite accuracy and ethics questions. Such inferences can cause harm regardless of how accurate they are. If they are accurate, they can make marginalized populations more legible and, therefore, vulnerable to abuses of state or social power. If they are inaccurate, the wrong people can be punished, denied benefits, or stripped of opportunities.

4. Race and Ethnicity

Race and ethnicity are inferable from many types of personal data, such as location and photos. For example, the Consumer Financial Protection Bureau (CFPB) was able to infer race and ethnicity from a combination of geography and surname information in mortgage applications.[116] The Equal Credit Opportunity Act prohibits creditors from finding out about race, ethnicity, or gender, but the CFPB wanted to obtain this data to identify potential discrimination.[117] The CFPB used name, geography, and general demographic information to infer people's race and ethnicity.[118]

Other types of data can give rise to inferences about race and ethnicity. In their analysis of Facebook likes, researchers correctly identified people's race 95% of the time.[119] In one study, researchers reliably inferred the race of patients based on doctor's notes where all explicit indications of race were removed.[120] The algorithm the researchers developed discerned patterns based on types of health conditions as well as troubling patterns of caregiver notes, which referred to Black patients more negatively than to other patients, labeling them as "difficult" or "demanding."[121]

---

[115] Sam Levin, *LGBT Groups Denounce 'Dangerous' AI That Uses Your Face to Guess Sexuality*, GUARDIAN (Sept. 8, 2017), https://www.theguardian.com/world/2017/sep/08/ai-gay-gaydar-algorithm-facial-recognition-criticism-stanford [https://perma.cc/35C5-Y8YS]; *Row over AI That 'Identifies Gay Faces,'* BBC NEWS (Sept. 11, 2017), https://www.bbc.co.uk/news/technology-41188560.amp [https://perma.cc/7U24-AH4F].

[116] CONSUMER FIN PROT. BUREAU, USING PUBLICLY AVAILABLE INFORMATION TO PROXY FOR UNIDENTIFIED RACE & ETHNICITY: A METHODOLOGY & ASSESSMENT 3 (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf [https://perma.cc/5CY4-NK3L].

[117] *Id.* at 4.

[118] *Id.* at 23.

[119] Kosinski et al., *supra* note 91, at 5803.

[120] Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Santeio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh & Marzyeh Ghassemi, *Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations*, 2022 PROC. AAAI/ACM CONF. ON AI, ETHICS, & SOC. 14.

[121] *Id.* at 15.

* * *

As the examples above demonstrate, inferences about sensitive data can readily be made from nonsensitive data. In many cases, such inferences can be made from innocuous and mundane data. Indeed, when fed into the right algorithm, nearly any data about a person can be used to make inferences about sensitive data.

C. The Dynamic Evolution of Inference

Today it is possible to make inferences about sensitive data from so many different types of nonsensitive data that sensitive data threatens to expand and engulf everything. Tomorrow, even more will be possible. As Tal Zarsky aptly notes, "over time and given Big Data analysis, 'special categories' mushroom in size."[122]

Thus, if it is not checkmate today, checkmate is just a few moves away, and there is no escape. The ability of algorithms to make inferences is developing at a staggering velocity. Labeling data as nonsensitive today might not hold for very long as machine learning algorithms discover new inference that can be made.

The implications of this conclusion are profound. Many organizations are violating the GDPR and other laws by not treating much personal data as sensitive.[123] In a dramatic upheaval, the rules for sensitive data would essentially become the main rules for processing most personal data while the rules for personal data would become a narrow or nonexistent exception. To comply, organizations might need to treat all personal data as sensitive, as it would be difficult to know for sure if data was not sensitive or would not become sensitive in the future.

D. Algorithms and Human Blind Spots

In practice, most attempts to identify sensitive data are rather crude and are merely based on human intuition and common sense. Privacy laws do not require organizations to examine the vast body of literature about what

---

[122] Zarsky, *supra* note 88, at 1013.

[123] Wachter and Mittelstadt note that some commentators contend that in order for personal data to be deemed sensitive, "the classification of data as sensitive depends on the stated purpose of processing. Data controllers must have the intention of inferring sensitive information from a selection of data for it to be classified as sensitive." Wachter & Mittelstadt, *supra* note 82, at 565. Wachter and Mittelstadt reject this view, noting that the Article 29 Working Party has taken a view that does not require intention. *Id.* at 565–66. Wachter and Mittelstadt ultimately conclude that "the classification of data, which indirectly reveals or can be used to infer sensitive information, is not so straightforward. The necessity of intentionality and reliability are a point of disagreement among commentators." *Id.* at 568.

inferences are possible. Instead, the laws seem to assume that identifying sensitive data will be as easy as sorting apples and oranges.

Research shows that humans are less capable than computers in making inferences from nonsensitive data, and this reveals a troubling problem—humans have many blind spots. They cannot see what algorithms can infer.

Several studies involving algorithms examined how well humans could make inferences based on the same data fed to machines. The studies revealed that the computers are more accurate—and often by a large margin. Recall the study where an algorithm could determine the race of patients based on doctors' notes. A group of physicians reviewing the same notes were much less capable of correctly identifying race.[124]

Another study revealed that computers were better able than humans to make assessments of people's personalities: "[C]omputers' judgments of people's personalities based on their digital footprints are more accurate and valid than judgments made by their close others or acquaintances (friends, family, spouse, colleagues, etc.)."[125] Moreover, "computer-based personality judgments were better at predicting life outcomes and other behaviorally related traits than human judgments."[126]

The fact that algorithms perform significantly better than humans in making inferences about sensitive data means that people cannot readily determine with their own common sense or intuitions the likelihood of how readily inferences can be made. Often, decisions by policymakers and organizations about what data could give rise to inferences about sensitive data are made without much examination of the research literature or without undertaking any research. Such decisions are made in an unsophisticated manner. But the studies I discussed are numerous and significant enough to throw the existing practices into doubt. The studies show that there will often be a significant risk that inferences about sensitive data can be made even from quite basic and innocuous information.

We thus have known blind spots when it comes to inferences about sensitive data. Any collection, processing, combination, or disclosure of regular personal data could unexpectedly give rise to inferences about sensitive data. We are walking in a minefield, and every step is treacherous. This situation is likely to grow worse as algorithms grow more sophisticated with machine learning.

---

[124] Adam et al., *supra* note 120, at 12.

[125] Wu Youyou, Michal Kosinski & David Stillwell, *Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans*, 112 PNAS 1036, 1036 (2015).

[126] *Id.* at 1039.

In summary, with inferences, nearly all personal data should be considered sensitive. To the extent that some personal data is not sensitive, the landscape of what inferences are possible constantly changes, so it might become sensitive in the future as different algorithms are developed or as different types of data are combined. Humans cannot easily determine on their own whether personal data is not sensitive. The sensitive data approach is thus unworkable.

The sensitive data approach has not fully reckoned with the ability of modern technology to make inferences. When such a reckoning occurs, sensitive data cannot survive. There is no fix.

III. THE NATURE OF DATA IS THE WRONG FOCUS

Even without the problem of inferences, sensitive data is unworkable. Beyond the fact that inference makes it nearly impossible to demarcate a separate realm of sensitive data, there is a fundamental problem at the root of sensitive data and deeply entwined in privacy laws—the idea that the appropriate protection of personal data can be determined by looking at the nature of the data.

Demarcating categories of sensitive data emerges from the view that the law should focus on the nature of the data. But this focus leads to the creation of arbitrary categories where some scenarios are deemed more sensitive without a clear or consistent rationale. This issue with classification is magnified when attempting to draw lines between the categories. At a distance, without much scrutiny, these categories might appear to be clear, but when they are looked at more closely, they are extremely blurry—so much so that they are practically useless.

In this Part, I argue that personal data protections should not turn on anything inherent in the data's nature. Privacy law should stop focusing on the nature of personal data. The particular type of personal data does not indicate anything important when it comes to determining how to protect it.[127] Instead, it creates arbitrary distinctions that are inconsistent between jurisdictions and impractical to administer. What matters is the harm or risk from collecting, using, or transferring personal data in particular situations.

A. Arbitrary Classifications and Blurry Lines

Sensitive data is an attempt at simplification; it makes the assumption that the collection, use, and disclosure of certain types of data generally

[127] DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 69 (2008) ("No particular kind of information or matter . . . is inherently private. The problem with focusing on the nature of the information or matter involved is that often there are strong privacy interests in relatively innocuous information or matters.").

can be more harmful or problematic than that of other types. These generalizations are too imprecise to make the distinction worthwhile.

Deeming data as "sensitive" is essentially a shortcut. Instead of a contextual and nuanced case-by-case analysis of each situation, demarcating categories of sensitive data avoids an alternative approach that is multifactorial and challenging to apply.

However, sensitive data categories are arbitrarily chosen, and their scope is too broad. Not all data that falls into a sensitive data category is equally sensitive. The fact that a person has a health condition might be quite embarrassing or harmful, or it might not be so at all. There are many people who voluntarily reveal this information to the public. Certain types of conditions are easier to conceal than others. Different conditions carry different stigmas and have different implications. The allure of sensitive data is to avoid blurry lines and complex case-by-case analyses. Unfortunately, however, sensitive data fails to achieve this goal. It does not solve the blurriness; it just shifts it. Instead of blurriness case-by-case, blurriness exists around the boundaries of sensitive data categories.

1. Arbitrariness

As discussed in Section I.B.2, the recognition of which categories of data are sensitive is quite inconsistent across laws. Several privacy laws in the United States recognize geolocation data as sensitive, but the GDPR does not. The GDPR recognizes philosophical beliefs as sensitive, and most U.S. laws do not. The different laws recognizing different categories of data as sensitive presents a complex mishmash that is not readily workable for organizations operating at a global (or even national) scale.

It is not clear that these lists are based on common views, as it does not appear that the drafters of laws conduct any polling or attempt any analysis to understand what people consider to be sensitive data. For example, in one survey in the United Kingdom in 2007, people rated financial data as the most sensitive type of data, which is not even included in the list of sensitive data in the Directive or the GDPR.[128] Of course, the United Kingdom is no longer in the EU, so this matter is moot. But do people in countries that are still in the EU consider financial data to be sensitive?

Perhaps this question is the wrong one to ask. People's views might be ill-informed, they might not fully understand the risks regarding certain types of personal data, and their opinions might change based on the news. Looking to societal attitudes is not an ideal approach, and it is wise for laws to avoid doing this.

[128] McCullagh, *supra* note 17, at 196–97.

How, then, are policymakers to decide which types of personal data to designate as sensitive? There does not appear to be any particular rule of recognition for sensitive data. There are no discernable theories or unifying principles.

The most plausible candidate for a theory that one can wrest from the morass is that sensitive data is more likely than regular personal data to cause harm. However, the special categories often do not correlate well with the processing of personal data that has a high risk of causing harm. As law professor Paul Ohm notes, categories of sensitive information are "not being thoughtfully or rigorously generated."[129] The creation of sensitive data lists is based on an anecdotal, nonscientific approach.

Another problem is that such lists differ from jurisdiction to jurisdiction. Philosophy matters in California and the EU, but apparently, it is not as important in Colorado, Connecticut, Utah, or Virginia. Only a medical or health diagnosis counts as health data in the United States, but in the EU and other countries, health is broader. Why doesn't health data matter if there is not a diagnosis? As there is no overarching theory or set of principles to determine what data should be deemed sensitive, the categories are arbitrary.

2. Blurry Lines

One challenge with sensitive data is that the various categories are often very loosely defined, if at all. The borders of the categories are so blurry and vague that they often beg the question of what types of data are included or excluded.

To understand blurry lines, first consider health as a sensitive data category. What constitutes health data? Of course, medical diagnoses by doctors are health data. But what about internet searches for health conditions or data generated when joining an online support group for a particular condition? Is one's fitness data health data? What about one's nutritional intake—data about all the food a person eats? Data about a person's body temperature, sleeping habits, physical activity level, smoking habits, consumption of alcohol, and other drug use all relate to health. However, there are no clear criteria to determine whether this data should fall into the category of sensitive health data.

Health becomes even more complicated when mental health is involved. For people with depression, data about their emotional state, such as an indication they are happy or sad, can involve health. Even their overall level of social activity can be an indication of health, as social withdrawal can be a sign of depression.

---

[129] Ohm, *supra* note 88, at 1139.

Turning to another example, what constitutes a religious belief? Is atheism a religious belief or instead a philosophical or scientific one? Is liking *On the Origin of Species* by Charles Darwin on Facebook an expression of a religious, philosophical, or scientific belief? In contrast to religious beliefs, scientific beliefs are not included on many lists of sensitive data. But perhaps a scientific belief could be considered a philosophical belief, which is a category of sensitive data under many laws. Categorizing science as philosophy requires resolving a philosophical debate spanning millennia. For some, philosophy could extend to any form of nonreligious belief.

What is a philosophical belief? The answer, ironically, depends upon one's philosophical beliefs. Nearly anything can be a philosophical belief depending upon one's conception of what "philosophy" is—an age-old issue that still lacks a definitive answer today. Yet, somehow, an answer to this question must be conjured up to figure out which beliefs should be included as sensitive data.

Because the lines between religion, science, and philosophy are blurry, the law could perhaps try to protect all beliefs rather than try to sort them into categories. Beliefs, after all, are often inextricably interlocked and they cannot be neatly separated into little boxes. One's religious beliefs will be connected to one's scientific, political, and philosophical beliefs. It is difficult to discern where one type begins and another type ends. For example, how does one categorize data about the purchase of books by the Marquise de Sade? These books are about sex, philosophy, religion, and politics.

It is even more difficult to determine how broadly a "belief" or "opinion" should be interpreted. Beliefs could include everything that constitutes a person's worldview. Yet, one's worldview consists of more than logic but also emotions, ideas, and various bric-a-brac cobbled together from movies, TV, books, the internet, life experiences, and more. Nearly everything one reads, writes, watches, and listens to is influenced by one's beliefs and also shapes one's beliefs.

Neil Richards aptly argues that protecting the privacy of thought, belief, reading, and communication are all components of what he calls "intellectual privacy," which he defines as a "zone of protection that guards our ability to make up our minds freely."[130] Richards contends that freedom of thought and belief "is the precondition for other political and religious rights guaranteed

[130] NEIL RICHARDS, INTELLECTUAL PRIVACY: RETHINKING CIVIL LIBERTIES IN THE DIGITAL AGE 95 (2015).

by the Western tradition."[131] Richards points to a long line of notable philosophers who have hailed the importance of freedom of thought and belief, especially John Stuart Mill, who extolled the need to protect "absolute freedom of opinion and sentiment on all subjects, practical or speculative, scientific, moral, or theological."[132]

Perhaps, to protect freedom of opinion, sensitive data should include any data that is the product of a person's intellectual activity. This would address the challenge of trying to determine where a "belief" or "opinion" ends and where other ideas, thoughts, or other stirrings of the mind begin. And, with this broad interpretation, then so much can be included. Richards contends that a person's online activity should be protected as intellectual privacy; he likens online searches to "a kind of thinking."[133] So most online activity—communication, searches, browsing, and so on—would all fall somewhere in the vast, blurry space between sensitive data categories. Likewise for offline activity—nearly anything people read, say, and share can be linked to their beliefs.

It is doubtful that privacy law can somehow figure out clear and coherent conceptions of the types of beliefs that are protected or excluded. But even if it could be figured out, privacy law is eons away from doing so. The journey has not even begun. And this journey would be a complicated one, destroying the illusion of simplicity that helps support the sensitive data approach in the first place.

B. The Harmfulness of Nonsensitive Data

Sensitive data could be seen as an attempt to identify data that is at a higher risk of causing harm. It might be simplistic, but sometimes simple is better than perfect because ease of execution is a virtue. More complex approaches can fail more frequently, making them worse than a less perfect approach. Unfortunately, as discussed above, the sensitive data approach only appears to be simple. When scrutinized, the reality is that it is quite complex and unmanageable.

Nonsensitive data can be used in ways that cause harm—as much if not more than sensitive data. In this Section, I will first provide some examples of types of data that are not considered sensitive but that have the potential to cause serious harm. These notable omissions include metadata, addresses, personality types, photos, and social class data. Then, I explain how nonsensitive data can be used as a proxy to cause the same kinds of harm as

---

[131] *Id.* at 112.

[132] JOHN STUART MILL, ON LIBERTY 7 (Stefan Colli ed., 1989) (1859).

[133] RICHARDS, *supra* note 130, at 122.

sensitive data. The sensitive data approach thus underprotects excluded types of data.

1. Notable Omissions

a. Metadata

"Metadata" is a term to describe "data about data," a type of purportedly innocuous form of personal data about communications and the usage of digital products and services, such as data about tracking, properties, origin, file sizes and titles, creation date, and more.[134] Metadata is noncontent information. For example, phone numbers and call duration are considered to be metadata; the conversation during the call is content. Email headers are also deemed to be metadata because they consist of routing information; the email message itself is content.

U.S. law has long attempted to single out metadata for lesser protection than content information. Such an attempt has proven to be a fool's errand. President Barack Obama famously attempted to justify the National Security Agency's improper surveillance by downplaying the importance of metadata:

> [W]hat the intelligence community is doing is looking at phone numbers and durations of calls. They are not looking at people's names, and they're not looking at content. But by sifting through this so-called metadata, they may identify potential leads with respect to folks who might engage in terrorism.[135]

Under the Fourth Amendment, U.S. Supreme Court precedent and federal and state electronic surveillance statutes treated certain kinds of data as less important than other types of data. In particular, with regard to a telephone call or email, their contents are protected more stringently by the Fourth Amendment and electronic surveillance statutes than the metadata associated with them.[136] The term "sensitive data" has not generally been used in discussions of metadata, but essentially, the law is attempting to make a distinction between types of data based on their nature—the same thing that sensitive data provisions seek to do.

---

[134] Chiradeep BasuMallick, *What is Metadata? Definition, Types, Uses, and Examples*, SPICEWORKS (Oct. 20, 2022), https://www.spiceworks.com/tech/devops/articles/what-is-metadata/ [https://perma.cc/BN9D-8938].

[135] Remarks on Healthcare Reform and an Exchange with Reporters in San Jose, California, 1 PUB. PAPERS 542, 545 (June 7, 2013).

[136] *See generally* Daniel J. Solove, *Reconstructing Electronic Surveillance Law*, 72 GEO. WASH. L. REV. 1264 (2004) (describing how the Fourth Amendment and electronic surveillance law regulate email, telephone calls, and other technologies).

Originally, metadata involved phone numbers dialed. In *Smith v. Maryland*, the U.S. Supreme Court held that a pen register, which recorded phone numbers dialed, was not covered by the Fourth Amendment.[137]

With more modern communications, a debate arose over what types of data should be analogous to phone numbers dialed. In the USA PATRIOT Act, Congress settled on a confusing and contradictory approach. It expanded the definition of the Pen Register Act to cover not just phone numbers dialed but any "routing" information. But then it stated that the routing information shall not involve the contents of the communication.[138]

The simple distinction forged in the 1970s in *Smith* between phone numbers and the contents of a phone call does not map on well to modern technologies of digital communications and online activity. Consider an IP address—the unique number assigned to each computer connected to the internet. IP addresses are called "addresses" because they are simply indications of location, identifying which particular computers the information is from. But aggregating a list of IP addresses that a person visits can reveal how a person navigates the internet, which can show a lot about that person's life, interests, and activities.[139]

A URL is even more revealing because a URL indicates a particular page. A website has the same IP address for all its pages, but each page has a different URL. URLs thus provide a more granular portrait of how a person is engaging with the internet and what information that person is seeking and consuming.[140] Are IP addresses and URLs really routing information? This issue remains quite unclear.

The very attempt to distinguish between routing information and content information is faulty. Routing information can be very revealing. Content information can sometimes be revealing. What people might care most about is protecting the privacy of the people and organizations they deal with, not the specific things they say. Even phone numbers can be quite revealing when traced back to specific people that a person is communicating with.

The understanding of the U.S. Supreme Court in the 1970s is quaint and obsolete in light of today's data analytics. Various pieces of innocuous data can be combined and analyzed to reveal extensive information about a person. One study found that "telephone metadata is densely interconnected,

---

[137] Smith v. Maryland, 442 U.S. 735, 745–46 (1979).

[138] Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT Act) Act of 2001, Pub. L. No. 107-56, § 216(c)(2)(A), 115 Stat. 272, 290 (2001).

[139] Solove, *supra* note 136, at 1287.

[140] *Id.* at 1287–88.

susceptible to reidentification, and enables highly sensitive inferences."[141] The U.S. Supreme Court has recognized this point with regard to geolocation data. In *Carpenter v. United States*, the Court recognized that people have a reasonable expectation of privacy in their geolocation data of their movement in public.[142] The location tracking occurred in public, which the Court previously determined was not within a person's expectation of privacy. Indeed, the Court held in *United States v. Knotts* that a person has no reasonable expectation of privacy when driving in public and tracked through a device because the movements were "voluntarily conveyed to anyone who wanted to look."[143] But in *Carpenter*, the Court was concerned about the extensiveness of the data. It noted that the geolocation data involved a "detailed chronicle of a person's physical presence compiled every day, every moment, over several years."[144]

Ultimately, the lesson is that the *type* of data is not the main issue in the analysis; the *extensiveness* of the data and how it might be used to make inferences about a person's private life is what matters. Sensitivity does not inhere in the data itself.

b. Addresses

Addresses are rarely on lists of sensitive data, yet they can be quite harmful if disclosed—sometimes a matter of life and death. Some sensitive data lists include geolocation data, which would purportedly include addresses, yet the laws do not extend to addresses, just data tracking a person's movement and location.[145]

For most people, the addresses of their home and work are quite innocuous. But for others, they are a matter of grave safety. Victims of stalking or domestic violence might want to protect their home and work addresses to hide from their tormentors. In one famous case, *Remsburg v. Docusearch, Inc.*, a woman was murdered by a stalker who bought her work address from a personal data search company.[146] In another instance, actress Rebecca Shaeffer was murdered when a stalker obtained her address from

[141] Jonathan Mayer, Patrick Mutchler & John C. Mitchell, *Evaluating the Privacy Properties of Telephone Metadata*, 113 PNAS 5536, 5536 (2016).

[142] 138 S. Ct. 2206, 2217 (2018).

[143] 460 U.S. 276, 281–82 (1983).

[144] 138 S. Ct. at 2220.

[145] California, Utah, and Virginia include geolocation data as sensitive data. CAL. CIV. CODE § 1798.140(ae)(1)(C) (West 2023); UTAH CODE ANN. § 13-61-101(32)(a)(iii) (West 2023); VA. CODE ANN. § 59.1-575 (2023). Colorado does not. COLO. REV. STAT. § 6-1-1303(24) (2023).

[146] 816 A.2d 1001, 1005–06 (N.H. 2003).

California motor vehicle records, sparking the passage of the federal Driver's Privacy Protection Act.[147]

Abortion doctors and their families are often subjected to death threats.[148] Many have been murdered.[149] For them, their home addresses, work addresses, children's names and school addresses, as well as information about their vehicles can be among the most sensitive of data.

For many other reasons, people have been subjected to terrifying online harassment, including threats of violence, rape, and death.[150] An insidious form of intimidation is to "dox" people—to reveal data helpful in tracking them down—in order to facilitate others in attacking or threatening them.[151] In one series of incidents, harassers associated with an online harassment campaign known as "Gamergate" attacked female game developer Brianna Wu by doxing.[152] She felt it was unsafe to return to her home, and she lived in terror.[153] In addition to issuing threats, others use information from doxing to engage in a practice called "swatting," which involves falsely calling in a threat to the police or fire department to send out officials to an address.[154] Swatting has sometimes led to the deaths of victims.[155]

Judges have been attacked at their homes. Wisconsin Judge John Roemer was murdered in his home by a defendant who appeared before him

---

147 DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 147 (2004).

148 *See* Planned Parenthood v. Am. Coal. of Life Activists, 290 F.3d 1058, 1065 (9th Cir. 2002) (en banc).

149 David S. Cohen & Krysten Connon, *Strikethrough (Fatality): The Origins of Online Stalking of Abortion Providers*, SLATE (May 21, 2015, 3:38 PM), https://slate.com/news-and-politics/2015/05/neal-horsley-of-nuremberg-files-died-true-threats-case-reconsidered-by-supreme-court-in-elonis.html [https://perma.cc/XE6W-WJT2].

150 DANIELLE KEATS CITRON, THE FIGHT FOR PRIVACY: PROTECTING DIGNITY, IDENTITY, AND LOVE IN THE DIGITAL AGE (2022).

151 Ryan Goodrich, *What Is Doxing?*, TECH NEWS DAILY (Apr. 2, 2013, 6:34 PM), https://web.archive.org/web/20141029095609/http://www.technewsdaily.com/17590-what-is-doxing.html [https://perma.cc/47FD-TE68].

152 Josh Fruhlinger, *What Is Doxing? Weaponizing Personal Information*, CSO (Aug. 31, 2020), https://www.csoonline.com/article/3572910/what-is-doxing-weaponizing-personal-information.html [https://perma.cc/3MKF-6SQ4].

153 Keith Stuart, *Brianna Wu and the Human Cost of Gamergate: 'Every Woman I Know in the Industry Is Scared,'* GUARDIAN (Oct. 17, 2014, 2:02 PM), https://www.theguardian.com/technology/2014/oct/17/brianna-wu-gamergate-human-cost [https://perma.cc/PLQ5-46DD].

154 Fruhlinger, *supra* note 152.

155 *See, e.g.*, Maria Cramer, *A Grandfather Died in 'Swatting' over His Twitter Handle, Officials Say*, N.Y. TIMES (July 24, 2021), https://www.nytimes.com/2021/07/24/us/mark-herring-swatting-tennessee.html [https://perma.cc/655X-FH93] (reporting that a man died of a heart attack at gunpoint while being swatted).

in court.[156] A gunman went to the home of federal Judge Esther Salas and killed her son and wounded her husband.[157] After the tragic attack, Judge Salas stated: "We preside over cases and 50% of the time people are not happy with us . . . . If the death of my 20-year-old son and now of Judge Roemer doesn't say we need something done to protect this personally identifiable information, I don't know what will."[158] In late 2022, the U.S. Congress passed the Daniel Anderl Judicial Security and Privacy Act, named after Judge Salas's murdered son. The law places restrictions on the sale and disclosure of judges' home addresses.[159] Although Congress recognized the importance of protecting the home addresses of judges, there are countless other people who are in just as much peril and who lack comparable protections.

c. Personality Type

Personality type is not included as sensitive data in most laws, but it is deeply related to a person's identity and selfhood, and data about personality can be used to manipulate and discriminate. "Personality" is a contested term, as there is a "lack of consensus" about how to define it.[160] Professor Dan McAdams offers a broad definition: "[P]ersonality is a developing configuration of psychological individuality that expresses a person's recognizable uniqueness, wherein life stories are layered over salient goals and values, which are layered over dispositional traits."[161] When personality is translated into data about people, it is often in the form of a personality type, a set of traits that serves as a profile. The most commonly referenced personality traits are known as the "Big 5": openness, conscientiousness, extroversion, agreeableness, and neuroticism.[162] Another widely discussed

---

[156] Eric Levenson & Boris Sanchez, *Federal Judge Whose Son Was Killed Two Years Ago Calls for Greater Judicial Protections After Former Wisconsin Judge Killed*, CNN (June 5, 2022, 12:30 PM), https://www.cnn.com/2022/06/05/us/wisconsin-judge-killed-attack/index.html [https://perma.cc/UDU6-WSKN].

[157] *Id.*

[158] *Id.*

[159] *See* Daniel Anderl Judicial Security and Privacy Act of 2021, S. 2340, 117th Cong. § 4(d)(1)(A) (2021) (enacted); *see also* Nate Raymond, *Judicial Security Measure Included in U.S. House-Passed Defense Policy Bill*, REUTERS (Dec. 8, 2022, 4:14 PM), https://www.reuters.com/legal/government/judicial-security-measure-included-us-house-passed-defense-policy-bill-2022-12-08/ [https://perma.cc/HL25-2STX].

[160] Susan C. Cloninger, *Conceptual and Historical Perspective*, *in* THE CAMBRIDGE HANDBOOK OF PERSONALITY PSYCHOLOGY 13, 13 (Philip T. Carr & Gerald Mattews eds., 2d ed. 2020).

[161] DAN P. MCADAMS, THE ART AND SCIENCE OF PERSONALITY DEVELOPMENT 8 (2015).

[162] Annabelle G.Y. Lim, *Big Five Personality Traits: The 5-Factor Model of Personality*, SIMPLY PSYCH. (July 10, 2023), https://www.simplypsychology.org/big-five-personality.html [https://perma.cc/L7B4-6KF2]; Courtney E. Ackerman, *Big Five Personality Traits: The OCEAN Model Explained*,

set of personality traits is the "Dark Triad": narcissism, Machiavellianism, and psychopathy. These traits are linked to negative behaviors, such as lying as well as exploiting or hurting others.[163] Being identified as having one or more of these personality traits can result in being considered a toxic and potentially dangerous person.

Personality type affects behavior, intellectual interest, health, politics, and more. As Renaud Lambiotte and Michal Kosinski observe: "Research has shown that personality is correlated with many aspects of life, including job success, attractiveness, drug use, marital satisfaction, infidelity, and happiness."[164] Data about personality can enable companies to manipulate behavior or make impactful decisions about people's lives. As previously referenced, Cambridge Analytica used personality information to manipulate people on Facebook to vote for Donald Trump and Brexit. The CEO of Cambridge Analytica extolled the ability to "sub-segment people by personality and change the creative to resonate with individuals based on how they see the world."[165]

Personality type can influence people's decisions from voting to purchasing behavior. One study revealed that "deep-seated personality traits can be linked to voting in theoretically consistent ways, over and above basic socio-demographic characteristics."[166] This study, which involved people in Germany, Greece, Italy, Poland, and Spain, found that personality traits were more strongly correlated to voting behavior than gender, age, income, or educational level.[167] Other studies demonstrated that openness is highly correlated with liberal political views in Belgium, Germany, Italy, Poland,

---

POSITIVE PSYCH. (June 23, 2017), https://positivepsychology.com/big-five-personality-theory/ [https://perma.cc/7DJN-BMGX]. These traits have effects on nearly every aspect of a person's life. Briefly, *openness* involves creativity, tolerance, and exploration of new ideas or things; *conscientiousness* involves being organized and consistent rather than spontaneous; *extroversion* involves being outgoing and social; *agreeableness* involves being kind and compassionate; and *neuroticism* involves being anxious and nervous. Renaud Lambiotte & Michal Kosinski, *Tracking the Digital Footprints of Personality*, 102 PROCS. IEEE 1934 (2014).

[163] Mia Belle Frothingham, *Dark Triad Personality Traits*, SIMPLY PSYCH. (July 26, 2023), https://www.simplypsychology.org/dark-triad-personality.html [https://perma.cc/G3S4-NLQR].

[164] Lambiotte & Kosinski, *supra* note 162; *see also* Jennifer Golbeck, Cristina Robles, Michon Edmondson & Karen Turner, *Predicting Personality from Twitter*, 2011 IEEE INT'L CONF. ON PRIV., SEC., RISK & TR. & IEEE INT'L CONF. ON SOC. COMPUTING 149, 149–156 ("Relationships have been discovered between personality and psychological disorders, job performance and satisfaction, and even romantic success.").

[165] Christopher Graves & Sandra Matz, *What Marketers Should Know About Personality-Based Marketing*, HARV. BUS. REV. (May 2, 2018), https://hbr.org/2018/05/what-marketers-should-know-about-personality-based-marketing [https://perma.cc/375B-GWSU].

[166] Michele Vecchione, Harald Schoen, José Luis González Castro, Jan Cieciuch, Vassilis Pavlopoulos & Gian Vittorio Caprara, *Personality Correlates of Party Preference: The Big Five in Five Big European Countries*, 51 PERSONALITY & INDIVIDUAL DIFFERENCES 737 (2011).

[167] *Id.*

and the United States, whereas conscientiousness is correlated with conservative views.[168]

In one study involving 3.5 million people, researchers found that "matching the content of persuasive appeals to individuals' psychological characteristics significantly altered their behavior as measured by clicks and purchases."[169] As Christopher Graves and Professor Sandra Matz declare in the *Harvard Business Review*: "The scientific evidence is consistent and clear: one can increase the effectiveness of marketing messages and other types of persuasive communication by tailoring them to people's psychological profiles."[170]

Studies and practice show that a wide array of types of data can be used to make inferences about personality. In one study, a linguistic analysis on people's blog posts consistently associated linguistic choices with personality traits.[171] The results "revealed robust correlations between the Big Five traits and the frequency with which bloggers used different word categories."[172]

Language use corresponds to personality type. Researchers developed word clouds associated with each of the Big Five personality traits based on analyzing 14.3 million Facebook statuses of roughly 75,000 volunteers who took a personality test. Patterns emerged that could be used to predict personality.[173]

One's "digital footprints, such as Facebook profile, or mobile device logs, can be used to infer personality."[174] One study predicted Big Five personality traits based on smart phone data including people's music listening, app usage, communication activity, and overall phone usage.[175]

---

[168] *Id.*

[169] S.C. Matz, M. Kosinski, G. Nave & D.J. Stillwell, *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, 114 PNAS 12,714, 12,714 (2017).

[170] Graves & Matz, *supra* note 165.

[171] Tal Yarkoni, *Personality in 100,000 Words: A Large-Scale Analysis of Personality and Word Use Among Bloggers*, 44 J. RSCH. PERSONALITY 363, 371 (2010).

[172] *Id.* at 366.

[173] H. Andrew Schwartz, Johnannes C. Eichstaedt, Lukasz Dziurzynski, Margaret L. Kern, Martin E.P. Seligman, Lyle H. Ungar, Eduardo Blanco, Michal Kosinski & David Stillwell, *Toward Personality Insights from Language Exploration in Social Media*, 2013 AAAI SPRING SYMP. 72, 76–77.

[174] Lambiotte & Kosinski, *supra* note 162, at 1934.

[175] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullman, Heinrich Hussman, Bernd Bischl & Markus Bühner, *Predicting Personality from Patterns of Behavior Collected with Smartphones*, 117 PNAS 17,680, 17,683 (2020).

Accurate inferences about personality were also made based upon people's Facebook activity.[176] Personality was found to be predictable based on the number of people a Twitter user follows, the number of followers a Twitter user has, and the number of times a Twitter user is listed in other people's reading lists.[177]

Personality type is a major focal point for marketers and influencers who seek to shape people's behavior. Data about personality type is deeply entwined with many aspects of a person's life and can be used in powerful ways to manipulate people. It is a notable omission from sensitive data lists.

d. Photos

Photos can be used in significantly harmful ways. Perhaps because of how widely photos are used, they are rarely included on sensitive data lists. But photos can readily reveal sensitive data. Nevertheless, the GDPR attempts to finesse the challenge that photos pose by stating at Recital 51:

> The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.[178]

The GDPR recital seems to view the only sensitive data category for photos as biometric data, but photos can lead to inferences about race, ethnicity, religion, health, and much more. For example, photos of people wearing religious clothing or with particular hair styles, facial hair, or head coverings can give rise to inferences about religion.

Various health conditions have physical manifestations that can be captured in a photo. Photos can reveal signs of drug use and addiction in the eyes or body. In one study, researchers developed a machine learning algorithm to predict depression based on photos people posted on Instagram—even before those people were diagnosed with depression. The algorithm performed better than general practitioners.[179]

---

[176] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell & Thore Graepel, *Manifestations of User Personality in Website Choice and Behaviour on Online Social Networks*, 95 MACH. LEARNING 357, 375–77 (2014).

[177] Daniele Quercia, Michal Kosinski, David Stillwell & Jon Crowcroft, *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*, 2011 IEEE INT'L CONF. ON PRIV., SEC., RISK & TR., & IEEE INT'L CONF. ON SOC. COMPUTING 180, 183.

[178] GDPR, *supra* note 2, § 51.

[179] Andrew G. Reece & Christopher M. Danforth, *Instagram Photos Reveal Predictive Markers of Depression*, 6 EPJ DATA SCI., no. 15, 2017, at 1, 9.

Even politics are inferable from photos. In one study, human research subjects accurately differentiated Democrats and Republicans based solely on their faces.[180]

In some situations, photos can be harmful even without any inferences being made. Although nude photos are not included on sensitive data lists, the practice of circulating nude photos of people without consent leads to considerable harm.[181] As Professor Mary Anne Franks notes:

> In a matter of days, that image can dominate the first several pages of search engine results for the victim's name, as well as being emailed or otherwise exhibited to the victim's family, employers, coworkers, and peers. Victims are frequently threatened with sexual assault, stalked, harassed, fired from jobs, and forced to change schools. Some victims have committed suicide.[182]

These harms are far more devastating than the release of a doctor's notes about a person's broken toe (health data) or information that a person is a Hegelian rather than Kantian (philosophical beliefs).

With the ready availability of photos online and their lack of protection, so much data can be inferred about a person's beliefs, behavior, and personality. A photo really is worth a thousand words.

e. Social Class

Data about social class involves various socioeconomic factors such as one's education and wealth.[183] Although discrimination based on social class is rampant, privacy laws rarely classify social class data as sensitive. People who are poor are subjected to significant discrimination and disparate treatment, including being subjected to a greater amount of surveillance.[184]

---

[180] Nicholas O. Rule & Nalini Ambady, *Democrats and Republicans Can Be Differentiated from Their Faces*, 5 PLOS ONE, Jan. 2010, at 1, 3, 5–6.

[181] Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 350–54 (2014); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 1–12 (2014).

[182] Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. 1251, 1259 (2017).

[183] *See, e.g.*, SOCIOLOGY: UNDERSTANDING AND CHANGING THE SOCIAL WORLD 241 (Univ. of Minn. Libs. Publ'g 2016) (2010); *see* ROUTLEDGE ENCYCLOPEDIA OF INTERNATIONAL POLITICAL ECONOMY 166 (R.J. Barry Jones ed., 2001); *The Class Structure in the U.S.*, COURSE SIDEKICK, https://www.coursehero.com/study-guides/boundless-sociology/the-class-structure-in-the-u-s/ [https://perma.cc/JEF7-V96W].

[184] YESHIMABEIT MILNER & AMY TRAUB, DATA FOR BLACK LIVES & DEMOS, DATA CAPITALISM + ALGORITHMIC RACISM 16 (2021), https://www.demos.org/sites/default/files/2021-05/Demos_%20D4BL_Data_Capitalism_Algorithmic_Racism.pdf [https://perma.cc/Z4C2-M82F] (describing the vast quantity of data required when applying for government benefits, including healthcare and nutrition assistance); Malkia Devich-Cyril, *Defund Facial Recognition*, ATLANTIC (July 5, 2020), https://www.theatlantic.com/technology/archive/2020/07/defund-facial-recognition/613771/ [https://perma.cc/5B84-AJAD] (describing the harmful effects of facial recognition technology on Black communities).

For example, as Professor Khiara Bridges argues, "due to the moral construction of poverty, there is a presumption that poor mothers will not put privacy rights to good uses. Indeed, the moral construction of poverty asserts that the poor are behaviorally and/or ethically flawed."[185] Mary Anne Franks contends that "[f]or the less privileged members of society, surveillance does not simply mean inhibited Internet searches or decreased willingness to make online purchases; it can mean an entire existence under scrutiny, with every personal choice carrying a risk of bodily harm."[186]

Preventing discrimination is one of the main rationales for including sensitive data categories in privacy laws; social class thus seems like an arbitrary exclusion. Additionally, social class is correlated to certain categories of sensitive data, such as race and ethnicity and political opinions.[187] Personal data about people is deeply intertwined, making it difficult to draw neat and tidy lines around certain data and separate it from other data.

2. Proxies

The primary rationales for sensitive data are to protect against situations involving a high risk to fundamental rights and freedoms or to protect against discrimination. Yet, these harms can readily be carried out with nonsensitive data.

In many cases, nonsensitive data can be used as a proxy for sensitive data.[188] For example, a postal code could be used as a proxy for people of a certain race or religion. As Professors Solon Barocas and Andrew Selbst note, algorithms can lead unintentionally to discriminatory results by using "proxy variables for protected classes."[189]

Even when nonsensitive data is not deliberately used as a proxy for a type of sensitive data, the correlation between nonsensitive data and sensitive data could have a harmful effect. For example, machine learning models "that were less likely to recommend Black patients to high-risk care management programs, more likely to identify Black defendants as high risk, and less likely to approve Black mortgage applicants all did not explicitly

---

[185] *See generally* KHIARA M. BRIDGES, THE POVERTY OF PRIVACY RIGHTS 12 (2017). For a detailed account of the extensive surveillance of mothers on welfare, see JOHN GILLIOM, OVERSEERS OF THE POOR: SURVEILLANCE, RESISTANCE, AND THE LIMITS OF PRIVACY (2001).

[186] Mary Anne Franks, *Democratic Surveillance*, 30 HARV. J.L. & TECH. 425, 453 (2017).

[187] OSCAR H. GANDY JR., COMING TO TERMS WITH CHANCE: ENGAGING RATIONAL DISCRIMINATION AND CUMULATIVE DISADVANTAGE 98 (2009) ("The racial composition of neighborhoods continues to be a very powerful predictor of the socioeconomic trajectory of those communities.").

[188] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 692–93 (2016).

[189] *Id.* at 675.

use race as a variable in making their predictions."[190] Thus, even unintentionally and without any malice, algorithms that use nonsensitive data can still lead to the same harms that the sensitive data approach seeks to avoid. Sensitive data can be stripped out of records, yet discrimination can still occur. These situations are quite troubling because they can appear as more neutral when sensitive data (race and ethnicity) is removed.

Clickstream data is not included on sensitive data lists, yet it can often be used as a proxy for sensitive data. Clickstream data can reveal a lot about people's race, religion, political opinions, and philosophical beliefs, among other types of sensitive data. Organizations using clickstream data do not need to infer sensitive data from it; they can just use it to target messages towards people or to manipulate their behavior for the same reasons they might have used sensitive data. The same aims that sensitive data are used for can be achieved with clickstream data without triggering sensitive data provisions.

3. Underprotection and Expressive Problems

The sensitive data approach excludes many very important categories where the law ought to provide stronger protection to personal data, resulting in underprotection of age and gender from discrimination. It relegates these categories to less protection on an arbitrary basis, and expressively connotes that these situations are less worthy of protection.

Even if nonsensitive data is not used as a proxy for sensitive data, it can still lead to the same type of harm. For example, there are other forms of discrimination beyond race and religion. As Wachter and Mittelstadt note, "gender, age, information about a person's financial situation, geolocation and personal profiles are not considered sensitive data under Article 9 [of the GDPR], despite often serving as grounds for discrimination."[191] If the goal of sensitive data is to curtail discrimination, it has major gaps.

Sensitive data elevates some forms of antidiscrimination above others. It ironically discriminates against many forms of illegal discrimination, such as age and gender. When privacy laws protect against discrimination of race, ethnicity, and sexual orientation but not age or gender, this relegates these unprotected forms of discrimination to a less important status.

Algorithms can usher in new forms of discrimination based on characteristics they identify as salient.[192] These characteristics might not be traditionally invidious ones such as race, gender, or age; they might be rather random characteristics based on uncanny correlations. For example, if

[190] Adam et al., *supra* note 120, at 8.

[191] Wachter & Mittelstadt, *supra* note 82, at 561.

[192] Zarsky, *supra* note 88, at 1013.

algorithms determine that having big feet correlates to successful job performance, they might use this characteristic. The result is that new undesirable characteristics will emerge, and they could be used systematically to people's benefit or detriment. A new form of inequality might arise, where people will be discriminated against based on having certain characteristics that they might not be able to change. This new inequality might be more hidden because algorithms can be quite complex.

Ultimately, what the law chooses to protect and what it omits have expressive impact.[193] These laws are expressing that some harms are less worthy of protection than other harms. How poorly sensitive data tracks harm is quite problematic. Privacy laws frequently ignore severe harms and elevate trivial harms for heightened protection.

Sensitive data creates the fiction that the law is addressing privacy problems proportionately to the seriousness of the harm or risks they pose when the law is, in fact, failing miserably in doing so. As a result, the wrong categories are being given extra protection for the wrong reasons, with policymakers thinking that they are somehow providing better and stronger privacy protection by including sensitive data in laws.

4. Personal Data Is a Grand Tapestry

The sensitive data categories are artificial constructs that are too simplistic. Different types of personal data blend into each other. For example, personality and psychiatric illness overlap, so personality is closely related to mental health. One's religious, philosophical, and political views are certainly influenced by one's personality, and vice versa. One's social class and finances also shape a person's beliefs and attitudes.

Due to the interconnectedness of the data, trying to isolate tidy categories of data is an impossible task. The existing categories have boundaries that nearly dissolve when one looks closely at data about a person and how it interrelates.

Looking at the extensive research about personal data inferences leads to a broad conclusion: Personal data is deeply intertwined with people and with other personal data. It is a grand tapestry, and the threads can't readily be pulled apart. Attempts to define categories of sensitive data try in vain to tease out different threads from the tapestry, but the threads are interwoven. When the threads are pulled apart, the whole tapestry unravels.

[193] Janice Nadler, *Expressive Law, Social Norms, and Social Groups*, 42 LAW & SOC. INQUIRY 60, 61–65 (2017) (expressing that what laws choose to regulate can influence public opinion on the importance of issues); Citron & Solove, *supra* note 1, at 816, 826.

IV. FOCUSING ON HARM AND RISK

The sensitive data approach falters because it is centered on a conceptual mistake—it views the nature of the data as the primary factor for determining the appropriate level of protection. As I discussed above, the nature of the data tells us little of value. What matters most is the harm and risk posed by collecting, using, or transferring personal data.

Harm involves negative consequences from the collection, use, or transfer of personal data that affect individuals or society. *Risk* involves the likelihood and gravity of certain harms that have not yet occurred.

In this Part, I discuss why the law should focus on harm and risk rather than pre-defined categories of sensitive data. Privacy law should provide more stringent protections based on the harm or risk of harm arising out of certain types of situations involving the collection, use, or transfer of personal data.[194] I discuss how such an approach would work, and I then address concerns about the complexity of assessing harm and risk more situationally and conclude that this approach is actually more practical than the sensitive data approach.

A. Proportionate Protection

As discussed in Part I, the sensitive data approach appropriately recognizes that not all situations involving personal data are the same and should not all be protected in the same way.

The law should not protect data for its own sake. The law should protect data to prevent or redress harm. Personal data in isolation is not inherently harmful. It becomes harmful or creates a risk of harm when it is used (or is likely to be used) in certain ways.

A more proportional approach is preferable to the simplistic two-level approach of sensitive data. The level of protection should vary proportionately to the harm or risk of harm. Specific protections should be directed to specific harms.

Of course, not all harms are knowable when a statute is enacted. So, a broad provision addressing unreasonable risk or unwarranted harm should be in place to cover anything that can arise. Known harms should be addressed, such as discrimination, manipulation, emotional distress, and reputational damage, among other things.

---

[194] By data "use," I am referring broadly to data processing, storage, and activities that organizations undertake with it. I thus use the term "use" to encompass what is done with personal data. It is important to note that the use of data is not the same as the purpose of the use. Purpose is the stated intention for a particular use, but a use might not necessarily be consistent with a stated purpose. Additionally, a use might aim to achieve a benign purpose yet have malignant side effects that cause harm or a risk of harm.

Risk and harm are certainly part of many privacy laws, but their role is not large enough. For example, the GDPR sometimes takes a risk-based approach. Article 24 looks to risk in its mandate for appropriate technical and organizational measures to protect data.[195] In Article 25, risk is a factor in evaluating what measures are appropriate for data protection by design and default.[196] Article 32 looks to risk for appropriate security measures.[197] And, in Article 35, risk is a key factor in triggering a requirement to conduct data protection impact assessments.

Unfortunately, the GDPR does not focus sufficiently on harm and risk in other provisions. For example, the GDPR requires that organizations appoint a data protection officer when the "core activities" of an organization involves "regular and systematic monitoring of data subjects on a large scale" or "processing on a large scale of special categories of data pursuant to Article 9 [sensitive data] or personal data relating to criminal convictions and offences referred to in Article 10."[198] There are countless uses that cause harm or a high risk of harm that fall outside of this provision. For example, the GDPR focuses on the large-scale processing of sensitive data but as discussed in Section III.B.1, many other types of personal data can cause significant harm, such as metadata, photos, social class, and so on. Perhaps the GDPR tries to address harm and risk by triggering the DPO requirement on the processing of sensitive data, but as discussed above, sensitive data poorly correlates to harm and risk. The sensitive data approach includes far too many situations that are not high risk and omits far too many situations that are high risk. These instances of inclusion and exclusion both cause problems.

The GDPR makes the same mistake with sensitive data elsewhere. In the DPIA requirement, although the GDPR focuses on situations involving a "high risk to the rights and freedoms of natural persons," it then lists the processing of sensitive data as a per se instance of high risk.[199] Including sensitive data here causes more mischief than good, as it wrongly encourages organizations to focus too much on sensitive data and underappreciate instances where nonsensitive data is involved. These flaws aside, the GDPR at least is on the right track by looking to risk in several provisions.

Some laws focus more generally on harm and risk for the DPIA requirement. For example, the CCPA and several other U.S. state consumer privacy laws look to a "heightened risk of harm to consumers" as a trigger

[195] GDPR, *supra* note 2, art. 24.

[196] *Id.* art. 25.

[197] *Id.* art. 32.

[198] *Id.* art. 37.

[199] *Id.* art. 35.1, 35.3(b).

for a privacy risk assessment.[200] These laws, unfortunately, have an odd circularity to privacy risk assessment requirements. The assessments are purportedly undertaken to identify risks, yet the risk must be surmised prior to the assessment. One would need to do the assessment to determine whether one was required. Ultimately, the initial judgment that there is a high-risk situation is often made based on readily apparent risk.

A more thorough risk-based approach would involve assessing risk more broadly, as a routine practice. A risk assessment should not be limited just to high-risk situations. Moderate risk is still significant and should not be ignored. Thus, the GDPR and all privacy laws should require a DPIA or privacy risk assessment for all forms of processing personal data. Harms and risks cannot be effectively addressed or minimized unless they are identified. Currently, risk assessment is far too infrequent and cursory. Although attention is showered on sensitive data, many other instances of processing personal data are given inadequate scrutiny.

B. Harm and Risk Depend Upon the Situation

Harm and risk can rarely be determined outside of context. For example, consider personal data that identifies a person as being of a particular faith. Many privacy laws would deem this to be sensitive data. But without knowing how the data will be used, it is not clear what protections are appropriate.

If the data about a person's religion is confidential, then the law should protect its confidentiality by restricting disclosure, imposing strong duties of confidentiality, and protecting the confidential relationships where this data is created and shared. But in many cases, data about religion is not confidential. Suppose the person is a well-known religious leader. Protection of this data as confidential would be meaningless—and even contrary to the person's desires.

If the data was used to discriminate against the person because of their faith, then this use would be harmful. Confidentiality protection would not be helpful since the data is already widely known. Meaningful protection would need to focus on stopping the data from being used to discriminate.

The law should address harms no matter what type of personal data is used—whether it be data directly about the person's religion, data that is a proxy for the person's religion, or data completely independent of the person's religion but used for these problematic purposes.

---

[200] CAL. CIV. CODE § 1798.185(a)(15) (West 2020), *amended by* 2023 Cal. Legis. Serv. 567 (West); COLO. REV. STAT. § 6-1-1309(2)(a)-(c) (2023); VA. CODE ANN. § 59.1-580(A)(5) (2023).

As this example demonstrates, the law's protections cannot be one-size-fits-all, as the particular harms and risks are quite different. Not every problem is the same. Looking at the nature of the data itself fails to tell us how it should be protected.

Turning to another example, the harms and risks involved with certain matters are different depending upon whether the data involves the present or future. For example, predictions (inferences made about the future) can cause considerable harms that are different from inferences about the present, which can be verifiable. As Hideyuki Matsumi and I contend, the lack of verifiability of predictions creates due process problems that are different from the use of nonpredictive data in decision-making. Mechanisms to ensure accuracy of data in privacy laws are ill-suited to protecting people against predictions involving forecasting the future.[201] Consider, for example, decisions made based on data about a person's past criminal convictions versus a prediction of a person's future crimes. The law should treat past versus predictive criminal data differently, as the latter creates risks to the presumption of innocence and other important societal values.

Treating all situations as equal often provides inadequate protections to high-risk situations. Another problem is treating low-risk situations with too many restrictions. Cumbersome and unnecessary restrictions trivialize privacy rules, making people perceive them as silly inconveniences and annoyances.

If privacy laws fail to focus on harm and risk, then they can perversely impede beneficial uses of data. Sensitive data provisions can be particularly stifling because they are restrictive. For example, Dominique Leipzig, Arsen Kourinian, David Biderman, and Tommy Tobin argue that because sensitive data includes data about race, restrictions on such information "threatens the ability of marginalized groups to access digital content."[202] They argue that "even though businesses may collect and share sensitive personal information for reasons beneficial for underrepresented communities, they may make a financial decision to stop doing so to avoid creating new compliance obligations implicated by collecting and disclosing sensitive information."[203] Advertisements targeted to certain racial groups might

---

[201] Hideyuki Matsumi & Daniel J. Solove, *The Prediction Society: Algorithms and the Problems of Forecasting the Future* 2, 60 (GWU Legal Stud., Rsch Paper No. 2023-58; GWU L. Sch. Pub. L., Rsch. Paper No. 2023-58, 2023), https://ssrn.com/abstract=4453869 [https://perma.cc/2Y7W-CMX6]; *see also* Matsumi, *supra* note 69, at 198–201.

[202] Dominique Shelton Leipzig, Arsen Kourinian, David Biderman & Tommy Tobin, *Ambiguity in CPRA Imperils Content Intended for Underrepresented Communities*, INT'L ASS'N OF PRIV. PROS. (Feb. 17, 2021), https://iapp.org/news/a/ambiguity-over-california-privacy-law-imperils-content-intended-for-underrepresented-communities/ [https://perma.cc/579J-4T7E].

[203] *Id.*

become challenging because race is sensitive data. Certain ads might be considered beneficial to certain racial groups, such as an ad promoting a diversity initiative. They argue that "online publishers may avoid creating, selling and/or using audience segments composed of individuals interested in issues impacting people of color and other historically underrepresented groups."[204] Additionally, it can "stifle speech related to identity and ideologies and hinder the publication of content related to social justice issues."[205]

The problem that Leipzig and her coauthors identify is caused by a failure of the law to look at harm and risk. They point to ways that data about race and ethnicity can be used positively in furtherance of inclusion and civil rights. Of course, such data can be used in bad ways too. The data itself in the abstract is not bad or good.

As some scholars note, excluding race, gender, or other characteristics from algorithmic decision-making does not always generate better results than when such characteristics are used. Professors Julian Nyarko, Sharad Goel, and Roseanna Sommers note that when criminal recidivism risk is assessed without accounting for gender (when data about males and females is considered together), the result is "an overestimation of the risk that female defendants will recidivate."[206] To better assess risk of female recidivism, the data of only females should be used. Thus, gender-blind data can yield less accurate results. They also note that because of racial bias in policing, race-blind studies "can overstate recidivism risk for Black individuals relative to white individuals. A similar phenomenon could, in theory, lead to higher auto insurance rates for Black and Hispanic drivers."[207] The use of data about race can help algorithms to correct for bias. Thus, the authors conclude, "avoiding the use of protected characteristics through the use of blind algorithms can, in some instances, lead to worse outcomes for members of a historically disadvantaged group."[208] Recall also the study by the CFPB discussed earlier. The CFPB needed data about race to study discrimination in mortgage applications. Because it was barred from doing so by other laws, it resorted to proxy data.[209]

Sensitive data provisions do not ban the use of race or other types of sensitive data, but they can be a strong deterrent to the collection and

---

[204] *Id.*

[205] *Id.*

[206] Julian Nyarko, Sharad Goel & Roseanna Sommers, *Breaking Taboos in Fair Machine Learning: An Experimental Study*, 2021 EAAMO '21, Oct. 2021, at 1, 2.

[207] *Id.*

[208] *Id.*

[209] *See supra* Section II.B.4.

processing of sensitive data because of added difficulties and expense. When there are beneficial uses of such data, the processing of the data should be encouraged rather than deterred.

Thus, the focus should not be on *data*, but instead about harmful or risky *uses of data*. For example, when data about race is used for illegal discrimination, then the processing should be banned. When uses of such data create a risk of illegal discrimination, greater scrutiny, restrictions, and oversight should be employed to prevent the risk from materializing into an actual harm. But if race were used in the way that the CFPB study wanted to use it—to do research to help combat discrimination—then such a use should be allowed. The law's protection shouldn't be triggered mechanically based on data involving race (as in the sensitive data approach); instead, the protection should be triggered by the harm or risk from the use.

Focusing on harm and risk can help avoid the problem of privacy being used as a pretext. Privacy becomes a pretext when invoked to achieve other aims that are not desired or helpful to the people whose privacy is purportedly being protected. As Professor Rory Van Loo notes, companies are using privacy as a pretext to hinder competition, reduce accountability, or achieve other goals that are unfavorable to consumers.[210] The privacy of customer data can be weaponized by companies seeking to impede lawsuits, regulatory investigations, or independent researchers.[211]

Heightened protection of race and ethnicity can undermine policies supporting people of color. In 2003, an anti-affirmative action referendum in California, the Racial Privacy Initiative, proposed banning the collection of data about race or ethnicity in order to attack affirmative action policies. The referendum was ultimately voted down. Professor Anita Allen observes that the referendum used the privacy protection of race as a pretext for attacking policies that actually benefited racial groups. On the other hand, Allen notes, "[t]he risks of government racial classification are clear when considering recent experiences in Rwanda, Bosnia, and Iraq. In those countries, slaughter and genocide were facilitated by quick reference to group membership recorded on an identification card."[212] Allen's discussion of the use of data about race or ethnicity demonstrates why privacy laws should focus on the use. Data can be used for good or ill.

Treating all sensitive data as the same encourages using privacy protections as a pretext to achieve other aims. These aims can be to cover up government or corporate wrongdoing or, as in the case of the California

---

[210] *See* Rory Van Loo, *Privacy Pretexts*, 108 CORNELL L. REV. 1, 3–4 (2022).

[211] *See id.* at 33.

[212] ANITA L. ALLEN, UNPOPULAR PRIVACY: WHAT MUST WE HIDE? 145 (2011).

referendum, to impede policies that help the very people whose privacy is purportedly being protected.

C. The Challenge of Complexity

The law shies away from focusing on harm and risk most likely because they are complicated and nuanced whereas sensitive data appears to be simple. But as I have demonstrated, the sensitive data approach is not really simple; any simplicity is just an illusion.

Nevertheless, even critics of sensitive data have a difficult time breaking free from the sensitive data approach because focusing on harm and risk is a daunting task. In one of the earliest and most extensive articles about sensitive data, Paul Ohm notes that the sensitive data approach can be arbitrary and lead to underprotection or overprotection of data.[213] Nevertheless, he concludes that sensitive data is worth the costs because of its simplicity.

Ohm argues that simplicity is the most practical approach. He notes that privacy harms identified in the work of certain scholars he labels as "New Privacy Scholars" are unlikely to be recognized by policymakers because these harms "lack the salience of traditional harms and are thus easy to ignore or outweigh; are stated so abstractly as to be incommensurable to other interests like security or economic efficiency; and do not lend themselves to testing or falsifiability."[214] These "New Privacy Scholars" include Paul Schwartz, Julie Cohen, Priscilla Regan, Anita Allen, and myself. Ohm observes that policymakers are not ready to embrace these theories of harm.[215]

The privacy harms that I and others have advanced are not quite as ethereal and unprecedented as Ohm implies. In a recent article I wrote with Professor Danielle Citron, we set forth a wide array of privacy harms that have a basis in existing law and cases.[216] We note that courts and policymakers are inconsistent in their recognition of privacy harms and that they can often falter and adopt narrow simplistic notions of harms rather than the broader and more pluralistic harms that we identify.[217] But the harms we identify have a basis in precedent and are not far-fetched. In an earlier article about data breach harms, we noted how some courts quickly stated that the law did not recognize emotional distress alone as cognizable harm, ignoring more than a century of indisputable precedent from hundreds (if not

---

213 *See* Ohm, *supra* note 88, at 1146.

214 *Id.* at 1147.

215 *Id.*

216 *See* Citron & Solove, *supra* note 1, at 799.

217 *See id.* at 747–56.

thousands) of privacy tort cases that did recognize emotional distress alone as sufficient to establish harm.[218] But as time has progressed, more courts have been recognizing harm in data breach cases.[219]

Thus, despite initial reluctance, there has been considerable movement by courts and policymakers towards recognizing harm. I do not believe that a lack of legal knowledge or imagination by some courts or policymakers presents an accurate indication of where courts and policymakers will end up in the future. The landscape of privacy law is constantly evolving. Policies that were inconceivable a few years ago are now widely accepted without a shrug. For example, prior to 2018, U.S. law did not recognize a right to data portability, and only a few laws had a very limited right to delete. Then, starting with the CCPA in 2018, several states have included these rights in their laws.[220]

Ohm certainly is right to be concerned that policymakers will find it challenging to develop a regulatory approach based on harm and risk. But he is wrong when he argues that "[r]einvigorating and expanding sensitive information law serves as a good second best alternative."[221] Ohm attempts to fix sensitive data by developing a theory to make it less arbitrary. He identifies four factors for identifying data as sensitive: "the possibility of harm; probability of harm; presence of a confidential relationship; and whether the risk reflects majoritarian concerns."[222] He recommends a "threat modeling" approach to analyzing harm.[223] Ohm also argues to expand sensitive data to also include precise geolocation, biometric data, and metadata.

Ironically, most of Ohm's efforts to improve sensitive data involve bringing more consideration of harm and risk under sensitive data's umbrella. It is not clear, however, how doing this makes harm less complex. Ohm's instinct is to focus on harm and risk, but he cannot bring himself to let go of sensitive data because of his attraction to its false Siren call of simplicity. Ultimately, clinging to sensitive data will impede his threat-modeling approach, which would be far better on its own without the impediments of sensitive data.

---

[218] *See id.* at 739.

[219] *See* James Dempsey, *US Courts Mixed on Letting Data Breach Suits Go Forward*, IAPP PRIV. PERSPS. (Mar. 9, 2022), https://iapp.org/news/a/u-s-courts-mixed-on-letting-data-breach-suits-go-forward/ [https://perma.cc/D9Q6-EMZ8].

[220] Solove, *supra* note 3, at 983.

[221] Ohm, *supra* note 88, at 1149.

[222] *Id.* at 1161.

[223] *Id.* at 1171.

The sensitive data approach is too flawed conceptually to be fixed. It is not simpler than focusing on harm and risk. Ohm underappreciates the problems with sensitive data, and he also concedes too much in his attempt to be pragmatic about what policymakers will do.

But being pragmatic ultimately means pushing for policy that actually works. The most pragmatic approach is to be frank with policymakers that sensitive data is a sinking ship. Adding a threat-modeling approach to the current regime would be like new fancy sails: going faster won't prevent the sinking. Instead, the most pragmatic strategy is to recommend sailing another ship. Yes, the approach of focusing on harm and risk might be a more difficult ship to sail, but it is possible to sail this ship, whereas it is not possible to continue on with the sinking ship of sensitive data.

Ultimately, there is no escape from the hard work of figuring out how to assess harm and risk. Privacy is immensely complicated, and it is highly contextual.[224] The law can protect against harmful uses of data by focusing on types of situations and uses rather than types of data. Of course, the law must make some generalizations and can't address each situation differently. But the fact that there are areas of contention and blurriness should not be a deterrent, as the boundaries of sensitive data are even less clear. Regulation that oversimplifies is ineffective—and often counterproductive—because it merely sweeps complexity under the rug.

CONCLUSION

Sensitive data is a key component of the GDPR and comprehensive privacy laws around the world. Sensitive data is also gaining popularity in the United States, finding its way into the new wave of state consumer privacy laws. Unfortunately, the sensitive data approach is unworkable. Privacy law must change course now before the problem is replicated throughout the states. Although sensitive data promises simplicity and practicality, these promises are illusory.

The law must adopt a risk–harm approach to realize the full power of modern data analytics and prepare for the profound power of future technologies. Simple distinctions based on the type of data are no longer meaningful in an age of inference because nearly all personal data can be sensitive. Instead, a risk–harm approach will be equipped to handle the implications of modern algorithms and inference. Such an approach will proportionately tailor regulation to harms and risks that emerge from data's

[224] *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE 7 (2010); *see* SOLOVE, *supra* note 127, at 1.

use. Despite reluctance to proceed with this approach, the law has no other viable option.