2023

Data Is What Data Does: Regulating Use, Harm, and Risk Instead of Sensitive Data

Daniel J. Solove
George Washington University Law School, dsolove@law.gwu.edu

Recommended Citation
118 Northwestern University Law Review (Forthcoming)

Data Is What Data Does:
Regulating Use, Harm, and Risk Instead of Sensitive Data

by

Daniel J. Solove

118 NORTHWESTERN UNIVERSITY LAW REVIEW (forthcoming 2024)

Draft: February 24, 2023

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 (GDPR), 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, biometric data, 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 non-sensitive 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 non-sensitive data can give rise to inferences about sensitive data. In many privacy laws, data that can give 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. Personal data is harmful when its use causes harm or creates a risk of harm. It is not harmful if it is not used in a way to cause harm or risk of harm.

To be effective, privacy law must focus on use, 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 based on the use of personal data and proportionate to the harm and risk involved with those uses.

DATA IS WHAT DATA DOES: REGULATING USE, HARM, AND RISK INSTEAD OF SENSITIVE DATA

by Daniel J. Solove[1]

[1] John Marshall Harlan Research Professor of 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 Alicia Solow-Niederman, Paul Ohm, and Paul Schwartz 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 a set of special categories of personal data 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, biometric data, and genetic data.[2]

Originally appearing in European Union (EU) data protection law, sensitive data has found its way into the comprehensive privacy laws of countless countries.[3] Long a holdout, the United States joined the bandwagon starting in 2020, when several state consumer privacy laws began including sensitive data.[4] 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 nothing meaningful for regulation can be determined solely by looking at the data itself. Data is what data does. Personal data is harmful when its use causes harm or creates a risk of harm. It is not harmful if it is not used in a way to cause harm or risk of harm.

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, it is easy to use non-sensitive 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.

[2] *See infra* Part I.B.3.

[3] *See infra* Part I.B.1.

[4] *See infra* Part I.B.1.

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 non-sensitive data can give rise to inferences about sensitive data, many privacy laws correctly consider it to be sensitive data.[5] Indeed in our age of modern data analytics, it would be naïve to fail to account for inferences. The problem, however, is that the rabbit hole goes all the way to Wonderland. With Big Data and powerful machine learning algorithms, most data can give rise to inferences about sensitive data, so arguably, nearly all personal data can be sensitive, and the sensitive data categories can swallow up everything. Oddly, the laws just seem to hum along as if this problem didn't 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 law that have heightened protections for sensitive data.

This Article contends that privacy law requires a rethinking. To be effective, privacy law must focus on use, 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 based on the use of personal data and proportionate to the harm and risk involved with those uses.

Currently, privacy statutes do not focus sufficiently on use, harm, and risk. Use is complicated and multifarious. Privacy harm and risk are issues that judges and policymakers have struggled over, especially in the United States.[6] Regulating based on use, 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. Complicated approaches are often challenging to execute and involve many stumbles along the way. Even if imperfect, a simple approach might be better than a complicated one. But the sensitive data approach only appears to be simple on the surface. When examined more deeply, the sensitive data approach is not as easy as it seems. Its simplicity is just an illusion; the complexity is still there. One can only pretend for so long that the elephant is not in the room.

The sensitive data approach might be defended as roughly tracking harm and risk, but the correlation is too weak to be useful. As I contend in this Article, sensitive data is not more harmful than non-sensitive data. It is the use of the data that matters.

---

[5] *See infra* Part II.A.

[6] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms, ,* 101 B.U. L. Rev. 793 (2022); Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data Breach Harms,* 96 Tex. L. Rev. 737 (2018).

This Article proceeds in several parts. Part I provides background about personal data and sensitive data. Part II examines the challenges that inference raises 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 use, 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.[7] Once a privacy law is triggered, it typically requires a slate of protections, which typically consist of two types: individual rights and organizational duties.[8]

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 allows for treating these situations with more protections.

In this Part, I discuss how personal data and sensitive data are defined, as well how sensitive data provisions in privacy laws function.

A. PERSONAL DATA

All privacy laws define the type of personal data that they cover.[9] 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.

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."[10] Data is *identified* if it is linked to a person. Data is *identifiable* if it could be linked to a person even if it is not currently connected. The linkage could 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,

[7] General Data Protection Regulation (GDPR), art. 4(1).

[8] Daniel J. Solove, *The Limitations of Privacy Rights,* 98 N.D. L. Rev. __ (forthcoming 2023).

[9] 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 US, the laws use a multitude of different terms. Some examples include: HIPAA: protected health information (PHI), Federal Communications Act: consumer network proprietary information (CPNI), FERPA: education records, Privacy Act: personally identifiable information (PII), CCPA: personal information.

[10] General Data Protection Regulation (GDPR), art. 4(1).

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 is thus identifiable to individuals, and it therefore is considered personal data under privacy laws that include identifiability in their definitions of personal data.

Data that is not about specific people, such as 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 65, is also not personal data. If privacy laws were to regulate all data, the laws would regulate the encyclopedia and nearly every piece of information. The laws would be completely incoherent.

How the 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.[11] For example, personal information under the U.S. Video Privacy Protection Act is defined as "information which identifies a person."[12] 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.[13]

The problem with this approach is that it 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.[14]

Although the 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.'"[15] The identified individuals approach is starting to wane in the U.S., with many of the newer laws defining personal data as relating to an identified *or an identifiable person.*[16]

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. Data that can reasonably be used in combination with other data to identify a person is personal data, even if in isolation it cannot be linked to a specific individual. Any data that is associated with personal data becomes personal data too.

---

[11] 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 (2011).

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

[13] Schwartz & Solove, *PII Problem, supra* note X, at __.

[14] Latanya Sweeney, Simple Demographics Often Identify People Uniquely 1 (Carnegie Mellon Univ., Sch. of Computer Sci., Data Privacy Lab., Working Paper No. 3, 2000).

[15] Graham Greenleaf, *California's CCPA 2.0: Does the US Finally Have a Data Privacy Act?,* 168 Privacy Laws & Business Int'l Report 13 (Dec. 2020).

[16] CPRA, Virginia, Colorado.

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.

In many circumstances, combining non-identifiable pieces of data can identify an individual. 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.[17]

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. Whether data is identifiable is best understood as a spectrum of risk. The risk can evolve over time based on changing conditions.

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, but another search query (such as a person's own name) is identifiable. The identifiability of a search query depends upon the specific query.[18]

B. SENSITIVE DATA

Countless privacy laws around the world have heightened protection for sensitive data. Commentators refer to sensitive data as "a bedrock of modern data protection."[19] Long an outlier, the United States has lacked a recognition of sensitive data in its privacy laws. Recently, however, the United States has joined the bandwagon. Since 2020, several new U.S. privacy laws started to recognize sensitive data — in particular, consumer privacy laws enacted by the states.[20]

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

[17] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of* Anonymization, 57 UCLA L. REV. 1701, 1723–25 (2010) (explaining re-identification techniques that combine anonymized data sets with outside information to accurately re-identify individuals), *see also* Henry T. Greely, *The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks*, 8 ANN. REV. OF 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).

[18] Schwartz & Solove, *PII Problem, supra* note X, at 134.

[19] Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1587.

[20] *See infra* at I.B.1.

are proportionate to the sensitivity of the data without having just two levels of protection.[21] Most laws, in contrast, do not have this more granular proportional approach.

1. Rise of Recognition of Sensitive Data

Sensitive data initially appeared in early privacy laws in Sweden and Hesse, a German state, in the early 1970s.[22] In its influential Privacy Guidelines of 1980, the Organization for Economic Co0operation and Development (OECD) recognized sensitive data, but merely had a barebones account of it, without specifying how it was to be protected or what types of data should be deemed to be sensitive.[23] The Guidelines states that "it is probably not possible to define a set of data which are universally regarded as being sensitive."[24]

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.[25] These categories were non-exclusive.[26]

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.[27] 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."[28]

As the sensitive data approach was taking form, a debate arose over whether it should be an open or closed list. An open list allows for new categories of sensitive data to be added over time; a closed list limits the categories to those specified in the law, and no additional categories can be added unless

---

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

[22] Karen McCullagh, *Data Sensitivity: Proposals for Resolving the Conundrum,* 2 J. Int'l Com. L. & Tech. 190, 190 (2007).

[23] Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1587.

[24] Explanatory Memorandum to the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980) at ¶ 19(a).

[25] Council of Europe., Explanatory Report on No. 108 of the Council of Europe Treaty Series—Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, ¶ 9 (1981), https://rm.coe.int/16800ca434.

[26] Karen McCullagh, *Data Sensitivity: Proposals for Resolving the Conundrum*, 2 J. Int'l Com. L. & Tech. 190 (2007) (p.2 of PDF) ("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.").

[27] United Nations Guidelines for the Regulation of Computerized Data Files G.A. Res. 45/95, para. 5 (Dec. 14, 1990).

[28] United Nations Guidelines for the Regulation of Computerized Data Files G.A. Res. 45/95, para. 5 (Dec. 14, 1990).

the law is amended.

In 1980, the OECD Privacy Guidelines took an open list approach. The explanatory memo to the OECD Privacy Guidelines noted 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."[29] The memo further noted that "It 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."[30] The memo went on to say: "There are precedents in European legislation to this effect (race, religious beliefs, criminal records, for instance). On 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."[31] The memo then stated that "The Expert Group discussed a number of sensitivity criteria, such as the risk of discrimination, but has not found it possible to define any set of data which are universally regarded as sensitive."[32]

Reflecting the view of the OECD, renowned EU data protection jurist Spiros Simitis contended in 1999 that "any personal datum can, depending on the purpose or the circumstances of the processing be sensitive. All data must consequently be assessed against the background of the context that determines their use."[33] Simitis contended that sensitivity should be "no more than a mere alarm device. It signals that the rules normally applicable to the processing of personal data may not secure adequate protection. Its primary consequence is therefore to incite a reflection process the purpose of which is to locate the shortcomings of the existing regulations and to establish the improvements needed."[34]

In 1995, the EU Data Protection Directive mandated that all EU member nations provide heightened protections for sensitive data. In contrast to the OECD Privacy Guidelines, the Directive specified types of sensitive data: "Member States shall 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."[35] The general approach of the Directive was to set the default rule to prohibiting 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 recipe when enacting laws, so each member state had to enact protections for the sensitive data categories

[29] Explanatory Memorandum to the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980) at ¶ 45.

[30] Explanatory Memorandum to the OECD Privacy Guidelines, *supra* note X, at ¶ 50.

[31] Explanatory Memorandum to the OECD Privacy Guidelines, *supra* note X, at ¶ 50.

[32] Explanatory Memorandum to the OECD Privacy Guidelines, *supra* note X, at ¶ 51.

[33] Spiros Simitis, *Revisiting Sensitive Data,* 5 (1999), https://rm.coe.int/09000016806845af.

[34] Simitis, *Revisiting Sensitive Data, supra* note X, at 8.

[35] European Council Directive 95/46, art. 8.

mentioned by the Directive. The Directive did not require its list to be a closed list; it was just a minimum list. Some countries enacted laws with open-ended lists of sensitive data, others enacted laws with closed lists.[36]

Following the DU Data Protection Directive, the GDPR was enacted in 2016 and added several additional categories of sensitive to the list in the Directive, including genetic data, biometric data, and sexual orientation.[37] However, unlike the Directive, which was an open list, the GDPR is a closed list. Member states cannot recognize additional categories of sensitive data, which they could under the Directive.[38] 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."[39]

The 1995 Directive and the GDPR have 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 (which were governed by the Directive).[40] Most privacy laws around the world have sensitive data protections.

Beyond the EU, countries 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.[41] Some countries recognize any data that could lead to discrimination as sensitive.[42]

---

[36] McCullagh, *Data Sensitivity, supra* note X, at p.10 of the PDF.

[37] GDPR art. 9.

[38] Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1589.

[39] GDPR art. 9(4).

[40] *See Data Protection Laws of the World: Turkey*, DLA Piper (Jan 13, 2022), https://www.dlapiperdataprotection.com/system/modules/za.co.heliosdesign.dla.lotw.data_protection/functions/handbook.pdf?country-1=TR (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); Argentina's Personal Data Protection Act No. 25.326 § 2 (Arg.) (defining sensitive data nearly identically to the EU Directive.), South Korea's 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 (Feb. 15, 2022), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=UY (Uruguay's Data Protection Act defines sensitive personal data with nearly identical verbiage to the EU Directive.).

[41] 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 means personal information that, once leaked or illegally used, may easily cause harm to the dignity of natural persons grave harm to personal or property security." PIPL, § 2 art. 28. The law then lists some examples, but the list does not appear to be exclusive.

[42] *See Data Protection Laws of the World: Japan*, DLA PIPER (Jan. 1, 2022), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=JP (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. 24, 2022), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=CO (Colombia defines sensitive data 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 (Dec. 15, 2021), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=EC

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). Subsequently, several states passed consumer privacy laws inspired by the CCPA and CPRA all include sensitive data: the Colorado Privacy Act (CPA), Virginia Consumer Data Protection Act, Utah Consumer Privacy Act, and Connecticut Act Concerning Personal Data Privacy and Online Monitoring.[43]

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
- genetic data
- biometric data.[44]

In a 2019 analysis of the definitions of sensitive data from 112 countries, the most commonly recognized categories of sensitive data include the types defined by the GDPR.[45] 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).[46] Other commonly recognized types of sensitive data include credit information and identification numbers or documents.[47] Some countries define sensitive data as including data related to private life, personal habits, or private relationships.[48]

(including data whose improper processing may give rise to discrimination in the definition of sensitive data), *and* 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).

[43] *See* VA. CODE ANN. § 59.1-575, COLO. REV. STAT. § 6-1-1303(24), UTAH CODE ANN. § 13-61-101(32), Conn. Public Act No. 22-15 § 1(27).

[44] GDPR art. 9.

[45] K Royal, *Sensitive Data Chart* (Sept. 13, 2019), on file with author.

[46] K Royal, *Sensitive Data Chart* (Sept. 13, 2019), on file with author. The 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 art. 10.

[47] K Royal, *Sensitive Data Chart* (Sept. 13, 2019), on file with author.

[48] K Royal, *Sensitive Data Chart* (Sept. 13, 2019), on file with author.

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.[49] Turkey uniquely recognizes clothing as sensitive data, likely because of the possibility that clothing can give rise to inferences about religion.[50]

In the United States, all of the state consumer privacy laws passed thus far (California, Connecticut, Colorado, Virginia, and Utah) recognize the following categories of sensitive data:

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

Many laws (except Utah and California) also recognize personal data collected from a known child as sensitive data. The laws all recognize citizenship or immigration status except California. Colorado recognizes sex life in addition to sexual orientation, but it lacks recognition of precise geolocation.

The California Consumer Privacy Protection Act 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
- trade union membership
- contents of mail, email, and text messages, unless the business is the intended recipient of the communication

Unlike the GDPR, the U.S. state 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 like 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, making compliance with the laws quite challenging. 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

[49] K Royal, *Sensitive Data Chart* (Sept. 13, 2019), on file with author.

[50] *See Data Protection Laws of the World: Turkey*, DLA Piper (Jan. 13, 2022), https://www.dlapiperdataprotection.com/index.html?t=definitions&c=TR (The Turkish LPPD includes clothing in its definition of sensitive personal data.).

(most of which include sensitive data),[51] plus laws in different U.S. states, mapping which data is sensitive is a complicated task because the same type of data may be sensitive in some jurisdictions but not others. Sometimes, categories are similar yet slightly different, such as "health" data under the GDPR versus "mental or physical health diagnosis" under the Virginia CDPA.

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.

The primary way that sensitive data is protected is to require express consent to process it under more 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 data relating to him or her."[52]

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) works to result in consent taking on 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.[53] The GDPR specifies six lawful bases:

> (a) The data subject has given consent to the processing of his or her personal data for one or more specific purposes;
> (b) 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;
> (c) processing is necessary for compliance with a legal obligation to which the controller is subject;
> (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
> (e) 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;

[51] According to the United Nations Conference on Trade and Development, 137 of 194 countries have comprehensive data privacy laws. UNCTD *Data Protection and Privacy Legislation Worldwide,* https://unctad.org/page/data-protection-and-privacy-legislation-worldwide. According to Graham Greenleaf, a professor who tracks global privacy laws, there are 157 countries with comprehensive privacy laws as of mid-March 2022. *See* Graham Greenleaf, *Now 157 Countries: Twelve Data Privacy Laws in 2021/22*, 176 Privacy Laws & Business International Report 1 (2022).

[52] General Data Protection Regulation, art. 4(11).

[53] GDPR art. 6.

> (f) 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.[54]

All personal data can only be collected and processed for one 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 identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.[55]

Then, Article 9 lists ten exceptions — allowable legal bases to process sensitive data.[56] 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
- research or statistical purposes

The sensitive data approach under the GDPR essentially functions as a series of two hurdles, the first hurdle involving a legal basis to process personal data and a second hurdle involving 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.[57] In most cases, once the first hurdle is cleared, the second hurdle can also be cleared. If data is processed with consent, then

[54] GDPR art. 6.

[55] GDPR art. 9(1).

[56] GDPR art. 9(2).

[57] GDPR art. 6.

it will clear the first and second hurdles since consent is a legal basis to process personal data and sensitive data. Some of the basis 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.

In practice, the main difference is that the legitimate interest basis cannot be used for sensitive data. For many organizations, legitimate interests is their primary way to process data without consent because the other legal basis are fairly narrow. 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 obtaining express consent, which can be quite difficult, the main way to process is through the legitimate interests legal basis. Thus, in practice, a key protection of sensitive data is to force organizations to obtain consent before processing when they otherwise would have been able to process via legitimate interests.

Other special protections for sensitive data in the GDPR include requiring the appointment of a data protection officer (DPO) and the carrying out of a data protection impact assessment (DPIA) when processing a "large scale of special categories of data."[58]

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. But the GDPR essentially removed any meaningful difference between the consent required for personal versus sensitive data.[59] Additionally, the GDPR does not allow member states to add further protections to sensitive data except for genetic, biometric, and health data.[60]

Moving beyond the GDPR, other countries provide similar protections for sensitive data. In the U.S., 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."[61] 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."[62] Essentially, the CCPA provides an opt out, as sensitive data may be processed unless individuals object.

Unlike the CCPA, the Virginia Consumer Data Protection Act (VCDPA) and

[58] GDPR art. 6.
1(c) (DPO); GDPR art 35.3(b) (DPIA)
[59] Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1601-02.
[60] GDPR art 9(4); Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1589.
[61] CAL. CIV. CODE § 1798.121(d) (West 2018).
[62] CAL. CIV. CODE § 1798.121(a) (West 2018).

the Colorado Privacy Act (CPA) require express consent and a data protection impact assessment to process sensitive data.[63]

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 involves focusing on the type of personal data involved to distinguish situations that should be afforded heightened protection. Under this view, some personal data can be quite innocuous. But other personal data can be very revealing, embarrassing, or damaging to one's reputation. On at least the surface level, data that a person is wearing a blue shirt does not appear to be particularly harmful or revealing. It seems quite innocuous. Contrast this to the fact that a person has a fatal disease. 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.

Quinn and Malgieri observe that EU law sometimes uses an instrumental rationale for protecting sensitive data but other times views protecting sensitive data as an end in and of itself.[64] For instrumental rationales, sources indicate two that are predominant: (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."[65] The European Court of Justice has quoted this language when explaining why sensitive data is protected more stringently.[66]

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 — that 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."[67]

---

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

[64] Quinn and Malgieri, *Defining Sensitive Data, supra* note X, at 1585-87.

[65] GDPR recital 51.

[66] OT v. Vyriausioji tarnybinės etikos komisija, Case 184/20 (CJEU Aug. 1, 2022) at ¶ 51.

[67] Explanatory Report on No. 223 of the Council of Europe Treaty Series—Protocol Amending the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, Council of Europe (2018), https://rm.coe.int/cets-223-explanatory-report-to-the-protocol-amending-the-convention-fo/16808ac91a.

The United Nations Guidelines for the Regulation of Computerized Data Files in 1990 included heightened protection of sensitive data because it created a risk of "unlawful or arbitrary discrimination."[68]

* * *

Sensitive data is now fully established as a key element of many privacy laws. Even the U.S., 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 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. NEARLY ALL DATA IS SENSITIVE: SENSITIVE DATA AND INFERENCE

We live today in what Alicia Solow-Niederman aptly calls an "inference economy."[69] 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 personal data.[70] Inferences about sensitive data can readily be made from non-sensitive data.[71] 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 being politicized.

Under several major privacy laws, it is clearly established that inferences count for 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. In today's age of Big Data, personal data is readily aggregated with other pieces of personal data and fed into hungry algorithms that generate inferences about people.

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, as they use machine learning and are trained on vaster

[68] United Nations Guidelines for the Regulation of Computerized Data Files G.A. Res. 45/95, para. 5 (Dec. 14, 1990).

[69] Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 Nw. L. Rev. 357 (2022).

[70] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016).

[71] 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 (2018).

quantities of data, they will be able to make more accurate and surprising inferences.

Privacy laws often gloss over this problem, but this is not a minor glitch to be tweaked. It is a fatal blow. In our age of inference, nearly all data is sensitive data.

A. INFERENCES COUNT

In the EU and other countries, data that could be used to make inferences about a category of sensitive data are included in that category. Under the GDPR, data that in combination could give rise to inferences about sensitive data is also deemed to be sensitive.[72] 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."[73]

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."[74] Thus, health data "also includes data about the purchase of medical products, devices and services, when health status can be *inferred* from the data."[75] In another guideline, the Article 29 Working Party (the predecessor to the European Data Protection Board) 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."[76]

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.[77] The case involved a law in Lithuania that required people receiving public funds to submit a declaration of interest, which included information

[72] European Data Protection Board, Advice Paper on Special Categories of Data ("Sensitive Data"), at 6 (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 (nothing 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.").

[73] European Data Protection Board, Guidelines on Automated Individual Decision Making and Profiling for the Purposes of Regulation, at 15 (2016), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[74] European Data Protection Board, Guidelines on Automated Individual Decision Making and Profiling for the Purposes of Regulation, at 15 (2016), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[75] Article 29 Data Protection Working Party, Annex – Health Data in Apps and Devices, at 1 (2015), http://ec.europa.eu/justice/data-protection/article-29/documentation/otherdocument/files/2015/20150205_letter_art29wp_ec_health_data_after_plenary_annex_en.pdf.

[76] Article 29 Data Protection Working Party, Guidelines on Automated Individual Decision Making and Profiling for the Purposes of Regulation, at 15 (2016), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[77] OT v. Vyriausioji tarnybinės etikos komisija, Case 184/20 (CJEU Aug. 1, 2022).

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 that "the 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."[78] 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."[79] 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."[80] 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.[81]

One issue is whether the sensitivity of inference-producing data should be viewed objectively (based on the possibility of making inferences) or intentionally (based on the stated intentions of the data controller). Paul Quinn and Gianclaudio Malgieri note that EU law is inconsistent on this question and note difficulties in both approaches.[82] Sandra Wachter and Brent Mittelstadt note that guidance by the Article 29 Working Party has recognized that sensitivity is determined objectively rather than based on the intentions of those seeking to process data.[83]

In the United States, the California Consumer Privacy Act (CCPA) does not directly refer to sensitive data but recognizes inferences as a form of personal data.[84] the CCPA's broad definition of "personal information" includes "inferences 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."[85] This definition is one of the first to

[78] OT v. Vyriausioji tarnybinės etikos komisija, Case 184/20 (CJEU Aug. 1, 2022), at ¶ 117.

[79] *Id.* at ¶ 119.

[80] OT v. Vyriausioji tarnybinės etikos komisija, In Case C 184/20 (Aug. 1, 2022), at ¶ 128.

[81] OT v. Vyriausioji tarnybinės etikos komisija, In Case C 184/20 (Aug. 1, 2022), at ¶ 127.

[82] Paul Quinn and Gianclaudio Malgieri, *The Difficulty of Defining Sensitive Data—The Concept of Sensitive Data in the EU Data Protection Framework*, 22 German L.J. 1583, 1591-1609 (2021).

[83] 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 (2019).

[84] Jordan M. Blanke, *Protection for "Inferences Drawn:" A Comparison between the General Data Protection Rule and the California Consumer Privacy Act*, 2 GLOBAL PRIVACY L. REV. 81 (2020); Jordan Blanke, *The CCPA, 'Inferences Drawn,' and Federal Preemption*, 29 Richmond J. L. & Tech. __ (2022), https://ssrn.com/abstract=4219967.

[85] CCPA, Cal. Civ. Code § 1798.140(v)(1)(K).

explicitly encompass inferences.[86] 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.[64] 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.[87]

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

The Colorado Privacy Act (CPA) is even more explicit than the CCPA – it clearly includes inferences about sensitive data. The draft Colorado Privacy Act Rules define "sensitive data inferences" to include inferences about sensitive data made "alone or in combination with other data."[88]

Counting data giving 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 non-sensitive data could readily be used, thus allowing relatively easy navigation around any restrictions for sensitive data.

B. INFERENCE-A-RAMA

In today's world of sophisticated data analytics, it is quite easy to infer sensitive data from non-sensitive data.[89] In the section above, I argued that

---

[86] Jordan M. Blanke, *Protection for "Inferences Drawn:" A Comparison between the General Data Protection Rule and the California Consumer Privacy Act*, 2 GLOBAL PRIVACY L. REV. 81 , 92 (2020).

[87] California Office of the Attorney General, Opinion of Bob Bonta and Susan Duncan Lee, No. 20-303 (Mar. 10, 2022), at 13.

[88] Colorado Privacy Act Rules 4 CCR-904-3, Rule 2.02 (Sept. 29, 2022).

[89] 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, 561, 564 (2019) (noting the distinction between sensitive and non-sensitive data "is increasingly strained in the era of Big Data analytics"); Zarsky, *Incompatible, supra* note X, at 1013 ("Big Data potentially undermines the entire distinction between these categories."); Paul Quinn and Gianclaudio Malgieri, *The Difficulty of Defining Sensitive Data—The Concept of Sensitive Data in the EU Data Protection Framework*, 22 German L.J. 1583, 1590-91 (2021) (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 techniquest increase, we might expand the use and breath of categories of inferentially

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 in laws. 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 non-sensitive data include:

- Data about patterns of electricity use can be used infer that a person or household is an Orthodox Jew. Orthodox Jews do not use electricity on Saturdays.

- Data about food consumption can be used to infer religion, such as Muslims, Jews, Hindus, or other faiths that do not eat particular foods or eat particular foods on particular holidays.

- Data about food consumption can be used to infer health conditions, such as particular diets for particular conditions (such as gluten-free for celiac or sugar-free for diabetes).

- 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 non-sensitive data. An examination of 327 studies about inferences found that common inferable attributes included gender, age, politics, location, occupation, race and ethnicity, family and relationships, education, income, health, religion, sexual orientation, and social class.[90]

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 some types of sensitive data.

1. 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 was the winner.[91] This led to calls to boycott Lindell's pillows. Lindell took a "commonplace object" and "imbued it with a political ideology."[92] 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

sensitive information.").

[90] Joanne Hinds & Adam N. Joinson, What Demographic Attributes Do Our Digital Footprint Reveal? A Systematic Review, PLoS ONE 1 (2018).

[91] Elizabeth Chang, "MyPillow Boycott: How a Product Can Spark an Identity Crisis," Wash. Post., Feb. 12, 2021.

[92] *Id.*

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 can 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.

In a study involving the Facebook likes of nearly 60,000 people, researchers could infer political party 85% of the time.[93] In one study from the UK, researchers developed an algorithm that could correctly identify political leanings 86% of the time from Twitter activity.[94] 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.[95]

In one of the most notorious examples of making inferences about political opinions, Cambridge Analytica mined Facebook profile data to profile them and target political advertisements to them to vote for Donald Trump.[96] Additionally, the data was used to promote voting to leave in the Brexit referendum of 2016. Cambridge Analytica enticed people to take a personality quiz.[97] Cambridge Analytica then gained access to the personal data from friends of the people who took the quiz – about 87 million individuals.[98]

In one study, researchers were able to show that "information about the user's activity in non-political discussion forums alone can very accurately predict political ideology."[99] For example, the frequent use of the word "feel" was correlated with "economically left wing views."[100]

---

[93] Michal Kosinski et. al, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proceedings of the Nat. Acad. Of Sci. of the U.S. of Am. 5802, 5802-03 (2013).

[94] Antoine Boutet, *What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election,* PROCEEDINGS OF THE SIXTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, https://ojs.aaai.org/index.php/ICWSM/article/view/14283.

[95] Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer, *Predicting the Political Alignment of Twitter Users*, 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing 192 (2011).

[96] Matthew Rosenberg et. al, *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.

[97] *Id.*

[98] Daniel Susser, Beat Roessler, and Helen Nissenbaum, *Online Manipulation: Hidden Influences in a Digital World,* 4 Geo. L. Tech. Rev. 1, 10 (2019).

[99] Michael Kitchener, Nadini Anatharama, Simon Angus, and Paul A. Raschky, *Prediction Political Ideology from Digital Footprints* (May 2022), https://doi.org/10.48550/arXiv.2206.00397.

[100] *Id.*

One study went beyond a 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 orientations.[101]

2. Sexual Orientation

Sexual orientation can be readily inferred from social media activity. Researchers were able to infer sexuality 88% of the time by analyzing Facebook likes.[102] One's social network of friendships can also be used to infer sexual orientation.[103]

Accurate inferences about sexual orientation can be made based on photos of a person's face. In one study, an algorithm could correctly identify sexual orientation based on one facial image 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. The algorithm's accuracy increased with more photos; with five photos, the algorithm was correct for 91% of men and 83% of women.[104]

3. Health

According to the EDPB, any data that can give rise to an inference "about the actual health status or health risk of a person" constitutes health data.[105] If this is true, then nearly everything constitutes health data. Health data can readily be inferred from countless other types of non-sensitive data. So many things that people do, buy, and eat can affect health, as can gender, age, race, ethnicity, and location.

Several studies have shown how inferences about health can be made based on social media data. An analysis Facebook likes was able to infer substance abuse for more than two-thirds of the profiles analyzed.[106] Researchers were

---

[101] Daniel Preotuc-Pietro et al., *Beyond Binary Labels: Political Ideology Prediction of Twitter Users,* PROCEEDINGS OF 55TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 729, (2017).

[102] Michal Kosinski et. al, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proceedings of the Nat. Acad. Of Sci. of the U.S. of Am. 5802, 5802-03 (2013); *see also* Michal Kosinski et al. *Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines*. 70 American Psychologist 543 (2015), https://doi.org/10.1037/a0039210.

[103] Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook Friendships Expose Sexual Orientation*, 14 First Monday No. 10 (2009), http://doi.org/10.5210/fm.v14i10.2611.

[104] Yilun Wang & Michal Kosinski, *Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images*. 114 J. of Personality and Social Psychology 246 (2018).

[105] European Data Protection Board, Guidelines on Automated Individual Decision Making and Profiling for the Purposes of Regulation, at 15 (2016), https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[106] Michal Kosinski et. al, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proceedings of the Nat. Acad. Of Sci. of the U.S. of Am. 5802, 5802-03 (2013).

able to develop a model to identify people's mental illnesses, such as depression, bipolar, schizophrenia, and borderline personality disorder, based on their posts on the social media site Reddit.[107]

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."[108] 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 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. 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 actually was pregnant.[109]

Duhigg provided more detail about the story in his book, *The Power of Habit,* which was published in the same year.[110] 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."[111] Additionally, they bought vitamins, scent-free soap, and cotton balls.[112]

When the story broke, it became the prime example of the privacy problems of data analytics.[113] 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:

---

[107] Jina Kim, Jieon Lee, Eunil Park, & Jinyoung Han, *A Deep Learning Model for Detecting Mental Illness from User Content on Social Media*, 10 Scientific Reports 11846 (2020).

[108] Munmun De Choudhury, Scott Counts, Eric Horvitz, & Aaron Hoff, *Characterizing and Predicting Postpartum Depression from Shared Facebook Data*, Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing 626 (ACM, 2014).

[109] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times Magazine (Feb. 16, 2012).

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

[111] *Id.* at 194.

[112] *Id.*

[113] See Jordan M. Blanke, *Protection for "Inferences Drawn:" A Comparison between the General Data Protection Rule and the California Consumer Privacy Act*, 2 Global Privacy L. Rev. 81 (2020) (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/ ("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.").

> 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.[114]

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

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

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.[115] 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.[116] The CFPB was able to use name, geography, and general demographic information to infer people's race and ethnicity.[117]

Other types of data can give rise to inferences about race and ethnicity. In their analysis of Facebook likes, researchers were able to correctly identify people's race 95% of the time.[118] In one study, researchers were able to reliably infer the race of patients based on doctor's notes where all explicit indications of race were removed.[119] The algorithm the researchers developed was able to discern patterns based on types of health conditions as well as troubling patterns of caregiver notes, such as the fact that notes about Black patients referred to them more negatively, such as more often

---

[114] *Id.* at 209-10.

[115] Consumer Fin Prot. Bureau, *Using Publicly Available Information to Proxy for Unidentified Race & Ethnicity* 1,3 (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

[116] *Id.*

[117] *Id.* at 23.

[118] Michal Kosinski et. al, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 Proceedings of the Nat. Acad. Of Sci. of the U.S. of Am. 5802, 5802-03 (2013).

[119] Hammad Adam et. al, *Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations,* Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2022), https://doi.org/10.1145/3514094.3534203.

labeling them as "very difficult" or "very demanding."[120]

C. The Dynamic Evolution of Inference

Today it is possible to make inferences about sensitive data from so many different types of non-sensitive 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."[121]

Thus, if it's 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 non-sensitive 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.[122] In a dramatic upheaval, the rules for sensitive data would essentially become the main rules for processing most personal data, with the rules for personal data being a narrow or non-existent 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 will 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 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 non-sensitive data, and this reveals a troubling problem — humans have so many blind spots — they cannot see what algorithms can infer.

Several studies involving algorithms examined how well humans could make

---

[120] *Id.*

[121] Tal Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 Seton Hall L. Rev. 995, 1012 (2017).

[122] 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 X, at 565. Wachter & 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 & 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 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.[123]

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.)."[124] Moreover, "computer-based personality judgments were better at predicting life outcomes and other behaviorally related traits than human judgments."[125]

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 to high risk that inferences about sensitive data can be made even from quite basic and innocuous information. The studies show that the possibility that these inferences can be reliably made will surprise us.

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.

* * *

With inferences, probably most non-sensitive data is sensitive data, and nearly all personal data should be considered to be sensitive. To the extent that some personal data is not sensitive, the landscape of what inferences are possible constantly changes, so it might be sensitive in the future as different algorithms are developed or as different types of data are combined. Humans will not be able to easily determine on their own whether personal data is not sensitive. The sensitive data approach is thus unworkable.

---

[123] *See* Adam et al., *Write It Like You See It*. *supra* note X at X.

[124] Wu Youyou, Michal Kosinski, & David Stillwell, *Computer-Based Personality Judgments are More Accurate than those Made by Humans*, 112 Proceedings of the National Academy of Sciences 1036, 1036 (2015).

[125] Wu Youyou, Michal Kosinski, & David Stillwell, *Computer-Based Personality Judgments are More Accurate than those Made by Humans*, 112 Proceedings of the National Academy of Sciences 1036, 1039 (2015).

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

In this Part, I discuss why 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. In this Part, I argue that how personal data should be protected should not turn on anything inherent in its 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.[126] What matters is use.

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

1. Arbitrariness

The recognition of which categories of data are sensitive is quite inconsistent across laws. As discussed earlier, different laws worldwide recognize different types of data as sensitive. 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 US 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 data people consider to be sensitive. 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

[126] 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.").

data in the Directive or the GDPR.[127] 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 the 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; their opinions might change based on whatever cases are recently in the news. Looking to the societal attitudes is not an ideal approach, and it is wise for laws to avoid doing this.

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 to when processing personal data has a high risk of causing harm. As Paul Ohm notes, categories of sensitive information are "not being thoughtfully or rigorously generated."[128] 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 Virginia, Colorado, Connecticut, or Utah. Only a medical or health diagnosis counts for health data in the U.S. 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 determining what data should be deemed sensitive, the results 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.

What constitutes health data? Of course, medical diagnoses by doctors are health data. But what about Internet searches for health conditions? 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, smoking, and physical activity level relate to health. Data about a person's consumption of alcohol relates to health, as does drug use.

[127] McCullagh, *Data Sensitivity, supra* note X, at pp. 7-8 of the PDF.

[128] Paul Ohm, *Sensitive Information*, 88 S. Cal. L. Rev. 1125 (2015).

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.

Turning to another example, what constitutes religious beliefs? Is atheism a religious belief? A philosophical one? Or a scientific one? Is liking *On the Origin of the Species* by Charles Darwin a religious, philosophical, or scientific belief?

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. Is science philosophy? Some might say yes, but others might say no. What is philosophy? This issue is one that has been debated in philosophy for millennia. The answer, ironically, is that the issue of what is a philosophical belief is a matter of philosophy. For some, their philosophical views of what philosophy includes consists of a very broad tent, perhaps even extending to any form of non-religious belief.

Perhaps the law should 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. What about books by Foucault or Nietzsche?

When it comes to any type of belief or opinion, whether religious, philosophical, or political, how broadly should a "belief" or "opinion" be interpretated? Do beliefs include everything that constitutes a person's worldview? And if so, 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 shapes one's beliefs. One's tastes in intellectual consumption are often quite tethered to one's religion and politics. 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."[129] Richards contends that freedom of thought and belief "is the precondition for other political and religious rights guaranteed by the Western tradition."[130] Richards points to a long line of notable philosophers has hailed the importance of freedom of thought and

[129] NEIL M. RICHARDS, INTELLECTUAL PRIVACY: RETHINKING CIVIL LIBERTIES IN THE DIGITAL AGE 96 (2015).
[130] *Id.* at 112.

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."[131]

Perhaps, if read liberally, 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."[132] So most online activity — communication, searches, browsing, and so on — would all fall somewhere in the vast blurry orbs of sensitive data categories.

We need not stop with the Internet. Many of the things people read, say, and share can be linked to their beliefs. Food and alcohol consumption are not only linked to health but also to people's beliefs, as many religions have elaborate rules and rituals around food and alcohol. Clothing is also related to beliefs.

It is doubtful that this can somehow be figured out, with very clear and coherent conceptions of types of beliefs that are included 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.

* * *

Sensitive data categories can be too broad and simplistic. In other words, 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.

Deeming data as "sensitive" is essentially a shortcut and simplification. Instead of a contextual and nuanced case-by-case analysis of each situation, demarcating categories of sensitive data obviate the need to do the difficult analysis of each situation. Such simplification avoids an alternative approach that is too multifactorial and challenging to apply. The allure of sensitive data is to avoid blurry lines and complex case-by-case analysis. 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, the blurriness exists around the categories of sensitive data. At a distance, without much scrutiny, these categories might appear to be clear, but when they are looked at more closely, they are actually extremely burry — so much so that they are

[131] JOHN STUART MILL, ON LIBERTY 9 (Stefan Colli ed. 1989) (1859).

[132] RICHARDS, INTELLECTUAL PRIVACY, *supra* note X, at 122.

practically useless.

Sensitive data might be justified as a rough proxy for data that is harmful in some form — often used in ways that could adversely affect a person's life or that could be embarrassing or damaging to one's reputation. Under this view, although sensitive data is an imperfect proxy for harm, it might be a useful simplification. But this proxy is not good enough. Its imprecision is too significant, and its costs are too high. Sensitive data is quite unworkable in practice — it only works to the extent the problems are ignored.

B. The Harmfulness of Non-Sensitive 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 and more complex approaches can fail more frequently to make them worse than a less perfect approach. Unfortunately, as discussed above, sensitive data appears to be simple but is quite complex under the microscope.

Non-sensitive data can be used in ways to cause harm — as much if not more than sensitive data. In this Section, I will provide some examples. Additionally, non-sensitive data can be used as a proxy to cause the same kinds of harm as sensitive data.

1. Notable Omissions

(a) Metadata

"Metadata" is a term to describe a type of purportedly innocuous form of personal data involved with communications and Internet use. U.S. law has long attempted to single out metadata for lesser protection. Such an attempt has proven to be a fool's errand. President 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.[133]

Under the Fourth Amendment, the U.S. Supreme Court 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 stringently by the Fourth Amendment and electronic surveillance statutes than the metadata associated with them. The term "sensitive data" has not generally been used in discussions of metadata, but essentially, the law is attempting to make a

[133] Statement of the President (June 7, 2013), https://obamawhitehouse.archives.gov/the-press-office/2013/06/07/statement-president.

distinction between types of data based on its nature — the same thing that sensitive data provisions seek to do.

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.

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.

The simple distinctions forged in the 1970s in *Smith* are ill-suited to today's world. 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 — which particular computers 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.[134]

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 about how a person is engaging with the Internet and what information that person is seeking and consuming.[135] Are IP addresses and URLs really envelope information? This issue remains quite unclear.

The very attempt to distinguish between envelope and content information is faulty. Envelope information can be very revealing. Content information can sometimes not be revealing. What people might care most about is protecting the privacy of the people and organizations that 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. 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.[136] The location tracking occurred in public, which the Court previously determined was not within a person's expectation of privacy. Indeed, in the Court held in *United States v. Knotts* that a person has no reasonable expectation of privacy when tracked through a device when driving in public because the movements were "voluntarily conveyed

[134] Solove, *Reconstructing Electronic Surveillance Law,* 1287-88.

[135] Solove, *Reconstructing Electronic Surveillance Law,* 1287-88.

[136] Carpenter v. United States, 138 S.Ct. 2206 (2017).

to anyone who wanted to look."[137] 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."[138]

Ultimately, the lesson is that the *type* of data is not the main issue in the analysis and that the *extensiveness* of the data matters as well as how it might be used to make inferences about a person's private life. 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.[139]

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 staking or domestic violence might want to protect their home and work addresses to hide from their tormenters. In one famous case, *Remsburg v. Docusearch,* a woman was murdered by a stalker who bought her work address from a personal data search company.[140] In another instance, actress Rebecca Shaeffer was murdered when a stalker obtained her address from the Department of Motor Vehicles, sparking the passage of the federal Driver's Privacy Protection Act.[141]

Abortion doctors are often subjected to death threats for themselves and their families.[142] Many have been murdered.[143] 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 extensive online harassment, including threats of violence, rape, and death.[144] An insidious

---

[137] United States v. Knotts, 460 U. S. 276, 281 (1983).

[138] *Id.* at X.

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

[140] Remsburg v. Docusearch, 816 A.2d 1001 (N.H. 2003).

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

[142] Planned Parenthood v. American Coalition of Life Activists, 290 F.3d 1058, 1065 (9th Cir. 2002) (en banc).

[143] David S. Cohen & Krysten Connon, "Strikethrough (Fatality); The Origins of Online Stalking of Abortion Providers," Slate (May 21, 2015), https://slate.com/news-and-politics/2015/05/neal-horsley-of-nuremberg-files-died-true-threats-case-reconsidered-by-supreme-court-in-elonis.html.

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

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.[145] In one series of incidents, known as "Gamergate," harassers attacked female game developers through an extensive campaign of doxing.[146] Many of the women felt it was unsafe to return to their homes, and they lived in terror.[147] 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.[148] Swatting has sometimes lead to deaths of victims.[149]

Judges have been attacked at their homes. Wisconsin Judge John Roemer was murdered in his home by a defendant who appeared before him in court.[150] A gunman went to the home of federal judge Esther Salas and killed her son and wounded her husband.[151] 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."[152] In late 2022, 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.[153]

Although Congress recognized the importance of protecting the home addresses of judges, there are countless other people who are in just as much peril who lack comparable protections.

(c) Personality

Personality 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

---

[145] Ryan Goodrich, "What Is Doxing?," Tech News Daily (Apr. 2, 2013), https://web.archive.org/web/20141029095609/http://www.technewsdaily.com/17590-what-is-doxing.html.

[146] 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.

[147] Keith Stuart, "Brianna Wu and the Human Cost of Gamergate: 'Every Woman I Know in the Industry Is Scared'," The Guardian (Oct. 17, 2014).

[148] Fruhlinger, "Doxing," *supra* note X.

[149] By Maria Cramer, "A Grandfather Died in 'Swatting' Over His Twitter Handle, Officials Say," N.Y. Times (July 24, 2021).

[150] Eric Levenson and 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).

[151] *Id.*

[152] *Id.*

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

there is a "lack of consensus" about how to define it.[154] Professor Dan McAdams offers a broad definition: "*personality 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.*"[155]

Personality affects behavior, intellectual interest, health, politics, behavior, and more. As Renaud Lambiotte and Michal Kosinksi observe, "Research has shown that personality is correlated with many aspects of life, including job success, attractiveness, drug use, marital satisfaction, infidelity, and happiness."[156] Data about personality can enable companies to manipulate behavior or make impactful decisions about people's lives. Cambridge Analytica used personality 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."[157] As Christopher Graves and 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 profile."[158]

The common "Big 5" personality traits, which include openness, conscientiousness, extroversion, agreeableness, and neuroticism, are frequently studied.[159] 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.[160]

The "Dark Triad" of personality traits – narcissism, Machiavellianism, and psychopathy – are linked to negative behaviors, such as lying as well as

---

[154] 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).

[155] DAN P. MCADAMS, THE ART AND SCIENCE OF PERSONALITY DEVELOPMENT 8 (2016).

[156] Renaud Lambiotte & Michal Kosinski, *Tracking the Digital Footprints of Personality*. 102 Proceedings of the IEEE no. 12 (Dec. 2014); *see also* Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner, *Predicting Personality from Twitter*, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 149-156, https://doi/10.1109/PASSAT/SocialCom.2011.33 ("Relationships have been discovered between personality and psychological disorders, job performance and satisfaction, and even romantic success.").

[157] Christopher Graves & Sandra Matz, *What Marketers Should Know About Personality-Based Marketing*, Harvard Business Review (May 2, 2018).

[158] *Id.*

[159] Annabelle G.Y. Lim, *What Are the Big 5 Personality Traits?,* Simply Psychology (June 15, 2020), https://www.simplypsychology.org/big-five-personality.html; Courtney E. Ackerman, *Big Five Personality Traits: The OCEAN Model Explained,* Positive Psychology (June 23, 2017), https://positivepsychology.com/big-five-personality-theory/.

[160] Renaud Lambiotte & Michal Kosinski, *Tracking the Digital Footprints of Personality*. 102 Proceedings of the IEEE no. 12 (Dec. 2014).

exploiting or hurting others.[161] Being identified as having one or more of these personality traits can result as being considered a toxic and potentially dangerous person.

One study revealed that "deep-seated personality traits can be linked to voting in theoretically consistent ways, over and above basic socio-demographic characteristics."[162] This study, which involved people in Italy, Spain, Germany, Greece, and Poland, found that personality traits were more strongly correlated to voting behavior than gender, age, income, or educational level.[163] Other studies demonstrated that openness is highly correlated with liberal political views in Germany, Italy, Belgium, Poland, and the United States whereas conscientiousness is correlated with conservative views.[164] Traits such as honesty can also be inferred from Facebook profiles.[165]

Personality can be used to influence people's purchasing, web browsing, or their voting or other decisions. 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."[166]

Studies and practice have shown 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 posts on blogs was used to accurately identify personality traits.[167] The results "revealed robust correlations between the Big Five traits and the frequency with which bloggers used different word categories."[168]

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 messages of 75,000 volunteers who took a personality test. Patterns emerged that are usable to predict personality.[169]

---

[161] Mia Belle Frothingham, *Dark Triad Personality Traits,* Simply Psychology (Mar. 13, 2022), https://www.simplypsychology.org/dark-triad-personality.html.

[162] 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 and Individual Differences 737 (2011).

[163] *Id.*

[164] *Id.*

[165] Jeffrey A. Hall & Natalie Pennington, *Self-Monitoring, Honesty, and Cue Use on Facebook: The Relationship with User Extraversion and Conscientiousness*, 29 Computers in Human Behavior 1556 (2013).

[166] Sandra Matz, Michal Kosinski, Gideon Nave, and David Stillwell, *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, Proceedings of the National Academy of Sciences 114(48):201710966 (Nov. 2017).

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

[168] *Id.* at p.5.

[169] H. Andrew Schwartz et al., *Toward Personality Insights from Language Exploration in Social Media,* AAAI Spring Symposium (Mar. 15, 2013), https://www.researchgate.net/publication/283270498_Toward_Personality_Insights_from_Language_Exploration_in_Social_Media.

One's "digital footprints, such as Facebook profile, or mobile device logs, can be used to infer personality."[170] One study was able to predict Big Five personality traits based on smart phone data involving people's music listening, app usage, communication activity, and overall phone usage.[171]

Accurate inferences about personality were also able to be made based upon people's web browsing and Facebook activity.[172] Personality was predictable based on the number of people Twitter users followed, the number of followers a Twitter user has, and the number of times a Twitter user is listed in other people's reading lists.[173]

Personality is a major focal point for marketers and influencers who seek to shape people's behavior. Personality 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.[174]

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, health, religion, and much more. Photos of people wearing religious clothing or with particular hair styles, facial hair, or head coverings can give rise to inferences of religion. Photos can reveal signs of drug use and addiction in the eyes or body. Various health conditions have physical manifestations that can be captured in a photo.

In one study, researchers developed a machine learning algorithm to predict depression based on photos people posted on Instagram – even before

[170] Renaud Lambiotte & Michal Kosinski, *Tracking the Digital Footprints of Personality*. 102 Proceedings of the IEEE no. 12 (Dec. 2014).

[171] Clemens Stachl et al, *Predicting Personality from Patterns of Behavior Collected With Smartphones*, 117 Proceedings of the National Academy of Sciences 17680 (2020).

[172] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel, *Manifestations of User Personality in Website Choice and Behaviour on Online Social Networks*, 95 Machine Learning 357 (2014).

[173] Daniele Quercia, Michal Kosinski, David Stillwell, & Jon Crowcroft, *Our Twitter Profiles, Our Selves: Predicting Personality with Twitter*, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing 180 (2011).

[174] GDPR recital 51.

people were diagnosed with depression. The algorithm performed better than "general practitioners."[175]

Even politics can be inferable from photos. In one study, human research subjects were able to differentiate Democrats and Republicans based on their faces.[176]

Beyond commonly recognized types of sensitive data, photos reveal quite a lot of intimate details about a person. Researchers were able to make inferences about personality traits based on the photos that people liked on Flickr.[177] One study demonstrated that personality traits such as extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience can be inferred from Facebook profile pictures.[178] Personality types could be accurately predicted by analyzing people's tweets on Twitter.[179]

The use of photos can cause significant damage to people. Although nude photos are not included on sensitive data lists, the practice of circulating nude photos of people without consent leads to considerable harm.[180] As 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."[181] These harms are far more devastating than the release of 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 lack of protection, so much data can be inferred about a person's beliefs, behavior, and personality. A

---

[175] Andrew G Reece & Christopher M Danforth, *Instagram Photos Reveal Predictive Markers of Depression*, 6 EPJ Data Science, article 15 (2017).

[176] Nicholas O. Rule & Nalini Ambady, *Democrats and Republicans Can Be Differentiated from Their Faces,* PLoS ONE 5(1): e8733. (2010), https://doi.org/10.1371/journal.pone.0008733.

[177] Crisitina Segalin. Alessandro Perina, Marco Cristani, and Alessandro Vinciarelli, *The Pictures we Like Are Our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits*, 8 IEEE Transactions on Affective Computing 268 (2017).

[178] Cristina Segalin et al., *What Your Facebook Profile Picture Reveals about Your Personality.* Proceedings of the 2017 ACM on Multimedia Conference 460 (2017), https://dl.acm.org/doi/10.1145/3123266.3123331.

[179] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner, *Predicting Personality from Twitter*, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 149-156, https://doi/10.1109/PASSAT/SocialCom.2011.33.

[180] Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 Wake Forest Law Review 345, 350-54 (2014); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 1-22 (2014).

[181] Mary Anne Franks, *"Revenge Porn" Reform: a View From the Front Lines*, 69 Fla. L. Rev. 1251, 1259 (2017).

photo really is worth a thousand words.

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 non-sensitive data.

In many cases, non-sensitive data can be used as a proxy for sensitive data.[182] For example, a postal code could be used as a proxy for people of a certain race or religion. Other data can be used as a proxy for race if race could readily be inferred from that data. As Solon Barocas and Andrew Selbst note, algorithms can lead unintentionally to discriminatory results by using "proxy variables for protected classes."[183]

Even when non-sensitive data is not deliberately used as a proxy for a type of sensitive data, the correlation between non-sensitive 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 use race as a variable in making their predictions."[184] Thus, even unintentionally, and without any malice, algorithms that use non-sensitive data can still lead to the same harms that protecting sensitive data 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 because 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 clicksteam data do not need to infer sensitive data from it; they can just use it to target messages to 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. Expressive Problems and Underprotection

Even if non-sensitive 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 Middelstadt note,

---

[182] Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 692-93 (2016).

[183] Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 675 (2016).

[184] Hammad Adam et. al, *Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations,* Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2022), https://doi.org/10.1145/3514094.3534203.

"gender, age, information about a person's financial situation, geolocation and personal profiles are not considered sensitive data under Article 9, despite often serving as grounds for discrimination.[185] If the goal of sensitive data is to curtail discrimination, it has major gaps.

Sensitive data elevates some forms of anti-discrimination 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.[186] These characteristics might not be traditional invidious ones such as race, gender, or age; they might be rather random characteristics based on uncanny correlations. For example, if 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.[187] These laws are expressing that some harms are less worthy of protection than other harms. What is quite problematic is how poorly sensitive data tracks harm. Privacy laws frequently ignore severe harms and elevate trivial harms for heightened protection.

The problem with sensitive data goes beyond its unworkability. The sensitive data approach has negative effects because it excludes many very important situations where the law ought to provide stronger protection to personal data. It relegates these situations to less protection on an arbitrary basis, and expressively connotes that these situations are less worthy of 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 failing miserably in doing so. As a result, the wrong things 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.

---

[185] 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, 561 (2019).

[186] Zarsky, *Incompatible, supra* note X, at 1013.

[187] Janice Nadler, *Expressive Law, Social Norms, and Social* Groups, 42 J. AM. BAR FOUNDATION 60, 61–65 (2017) (expressing that what laws choose to regulate can influence public opinion on the importance of issues); Citron & Solove, *Privacy Harms, supra* note X, at X.

4. Personal Data Is a Grand Tapestry

The sensitive data categories are simplistic artificial constructs, and they are unfortunately too simplistic. Different types of personal data blend into each other. Personality disorders are a psychiatric diagnosis – personality is thus 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. The sensitive data categories focus mainly on certain traits and attitudes rather than behaviors. But behaviors and attitudes are quite often related.

Looking at the extensive research about inferences about personal data 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 threads are woven together with other threads. When one starts to pull apart the threads, the whole tapestry unravels.

IV. FOCUSING ON USE, HARM, AND RISK

The sensitive data approach falters because it is centered on a conceptual mistake — it views the nature of the data as significant for determining the appropriate level of protection. As I discussed above, the nature of the data tells us little of value. What matters is use, harm, and risk.

Use involves what is done with personal data, the activities that organizations undertake with it, and the purposes and effects of those activities. *Harm* involves negative consequence from the use 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 use, harm, and risk.

A. PROPORTIONATE PROTECTION

As discussed earlier, 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. Certain types of personal data are not inherently harmful. They become harmful or create a risk of harm when they are 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.

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.[188] In Article 25, risk is a factor in evaluating what measures are appropriate for data protection by design and default.[189] Article 32 looks to risk for appropriate security measures.[190] And, in Article 35, risk is a key factor in triggering a requirement to conduct data protection impact assessments (DPIAs).

Unfortunately, the GDPR does not focus sufficiently on use, harm, and risk in other provisions. For example, the GDPR requires that organizations appoint a data protection officer (DPO) 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 pursuant to Article 9 [sensitive data] or personal data relating to criminal convictions and offences referred to in Article 10."[191] There are countless uses that cause harm or a high risk of harm that fall outside of this provision. Perhaps the GDPR tries to address harm and risk through 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 sins 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.[192] 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 non-sensitive data is involved.

These flaws aside, the GDPR at least is on the right track by looking to risk in several provisions. Other laws focus less on risk. The California Consumer Privacy Act, for example, and most of the other U.S. state consumer privacy laws, primarily look to a "heightened risk of harm to consumers" as a trigger for a privacy risk assessment.[193]

---

[188] GDPR art. 24.

[189] GDPR art. 25.

[190] GDPR art. 32.

[191] GDPR art. 37.

[192] GDPR art. 35.1, 35.3(b).

[193] VDCPA, § 59.1-576(B), CPA, § 6-1-1309(2)(a)-(c).

There is generally an odd circularity to privacy risk assessment requirements trigged on high risk. 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.

B. USE CREATES HARMS AND RISKS

Some uses are good, some are bad, and some are in the middle. The more harm a use causes or the greater the risk of harm it poses, the more protection the law should provide.

The way data is protected depends upon use, harm, and risk. For example, consider personal data about a person's religion — that a person identifies 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 religious leader. Protection of this data as confidential would be meaningless — and even contrary to the person's desires, which might be to have this information widely known.

If the data were 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.

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

Turning to another example, the harms and risks involved with certain matters is 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 Yuki Matsumi contends, the lack of

verifiability of predictions creates due process problems that are different from the use of non-predictive data in decisonmaking. Mechanisms to ensure accuracy of data in privacy laws are ill-suited to protecting people against predictions involving forecasting the future.[194] Consider, for example, data about a person's criminal activity based on actual crimes committed versus a prediction of future crimes. The law should treat actual versus predictive criminal data differently, as the latter creates risks to the presumption of innocence and other important societal values.

The law should protect uses differently based on harm or the risk of harm involved. Treating all uses 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 use, 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."[195] 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."[196] Advertisements targeted to certain racial groups might 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." Additionally, it can "stifle speech related to identity and ideologies and hinder the publication of content related to social justice issues."[197]

The problem that Leipzig and her co-authors identify is caused by a failure of the law to look at use. 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 is not bad or good. The use is what matters.

As some scholars note, excising race, gender, or other characteristics from

---

[194] Hideyuki Matsumi, *Predictions and Privacy: Should There Be Rules About Using Personal Data to Forecast the Future?*, 48 Cumb. L. Rev. 149, 198-201 (2018).

[195] Dominique S. Leipzig, Arsen Kourinian, David Biderman & Tommy Tobin, Ambiguity in CPRA Imperils Content Intended for Underrepresented Communities, IAPP (Feb. 17, 2021), https://iapp.org/news/a/ambiguity-over-california-privacy-law-imperils-content-intended-for-underrepresented-communities/.

[196] *Id.*

[197] *Id.*

algorithmic decisionmaking does not always generate better results than when such characteristics are used. Julian Nyarko, Sharad Goel, and Roseanna Sommers note that when criminal recidivism risk is assessed without accounting for gender, the result is "an overestimation of the risk that female defendants will recidivate."[198] 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."[199] 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."[200] Recall 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.[201]

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 processing of this 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.

Focusing on use 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 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.[202] The privacy of customer data can be weaponized by companies seeking to impede lawsuits, regulatory investigations, and independent researchers.[203]

Heightened protection of race and ethnicity can undermine policies supporting people of color. In 2003, an anti-affirmative action referendum, 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. Anita Allen observes that the referendum used protecting the privacy of race as a pretext for attacking policies that actually benefited racial groups. On the other hand, Allen notes, "The risks of government racial classification are clear when considering recent experiences in Rwanda, Bosnia, and Iraq. In those countries,

---

[198] Julian Nyarko, Sharad Goel, Roseanna Sommers, *Breaking Taboos in Fair Machine Learning: An Experimental Study, Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* 1, 3 (EAAMO 2021), https://doi.org/10.1145/3465416.3483291.

[199] *Id.*

[200] *Id.*

[201] *See supra* Part II.B.4.

[202] Rory Van Loo, *Privacy Pretexts,* __ Cornell L. Rev. __ (forthcoming 2022) (manuscript on file with author), pp. 101-12.

[203] *Id.* at pp. 133-34

slaughter and genocide were facilitated by quick reference to group membership recorded on an identification card."[204] Allen's discussion of the use of data about race or ethnicity demonstrates why privacy laws should focus on use. Data can be used for good or ill.

Treating all sensitive data the same adds fuel to attempts to use 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 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 use, 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 use, 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 lead to underprotection or overprotection of data.[205] 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."[206] 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.[207]

To begin with a quibble, I would hardly characterize myself or the other scholars he mentions as "new" since we all started writing quite a long time ago. Turning to a more important point, 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 Danielle Citron, we set forth a wide array of privacy harms that have a basis in existing law and cases.[208] We note that courts and policymakers are inconsistent in their recognition of privacy

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

[205] Paul Ohm, *Sensitive Information*, 88 S. Cal. L. Rev. 1125, 1146 (2015).

[206] *Id.* at 1147.

[207] *Id.* at 1147.

[208] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms,* 101 B.U. L. Rev. 793 (2022).

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. 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 thousands) of privacy tort cases that did recognize emotional distress alone as sufficient to establish harm.[209]

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.[210]

Ohm certainly is right to be concerned that policymakers will find it challenging to develop a regulatory approach based on use, harm, and risk. He is wrong when he argues that "[r]einvigorating and expanding sensitive information law serves as a good second best alternative."[211] 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."[212] He recommends a "threat modeling" approach to analyzing harm.[213] 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 injecting considerations of harm into it. Instead of escaping from harm because of a concern regulators will resist it, Ohm urges that it be brought under sensitive data's umbrella. It is not clear, however, how doing this makes harm less complex. Moreover, Ohm invites harm in, but he omits use. Data itself is not harmful or risky. It is the *use* of the data that is.

The sensitive data approach is too flawed conceptually to be fixed. It is not simpler than focusing on use, risk, or harm. Ohm underappreciates the problems with sensitive data, and he also concedes too much in his attempt to be pragmatic about what policymakers will do. Being pragmatic ultimately means pushing for policy that actually works. The most pragmatic policy recommendation is to not mislead policymakers by telling them that it is okay to continue on the sinking ship of sensitive data and to give them new

[209] Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data Breach Harms,* 96 Tex. L. Rev. 737 (2018).

[210] Solove, *Limitations of Privacy Rights, supra* note X, at X.

[211] *Id.* at 1149.

[212] *Id.* at 1161.

[213] Paul Ohm, *Sensitive Information*, 88 S. Cal. L. Rev. 1125 (2015).

fancy sails so they can go faster. Instead, the most pragmatic strategy is to recommend that they find another ship. Yes, the approach of focusing on use, 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 sensitive data.

Focusing on use, harm, and risk is the only viable path, however difficult it may be. The sooner privacy law realizes this truth — as inconvenient as it may be — the better it will be for starting to journey down the right path.
Of course, not all use cases are clear. The harmfulness of many uses is in dispute. 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. Ultimately, there is no escape from the hard work of figuring out how to regulate certain uses. Privacy is immensely complicated, and it is highly contextual.[214] Regulation that oversimplifies is ineffective and often counterproductive because it does not reduce complexity but just shifts or hides it.

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 breed of state consumer privacy laws.

Unfortunately, sensitive data is unworkable. Although it promises simplicity and practicality, these benefits of sensitive data are illusory and only exist when the problems are ignored. The law must realize the full power of modern data analytics and the even more profound power of future technologies. Simple distinctions based on type of data are no longer meaningful in an age of inference. Nearly all personal data can be sensitive, and privacy law has not fully digested the implications of modern algorithms and inference.

There is no escape from focusing on use, harm, and risk. Despite reluctance to do so, the law has no other viable option.

[214] HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010); DANIEL J. SOLOVE, UNDERSTANDING PRIVACY (2007).