2013

Reconciling Personal Information in the United States and European Union

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

Paul M. Schwartz

RECONCILING PERSONAL INFORMATION IN THE UNITED STATES AND EUROPEAN UNION

By

Paul M. Schwartz[*]
Daniel J. Solove[**]

May 3, 2013

[*] Professor of Law, University of California, Berkeley, School of Law; Director, Berkeley Center for Law & Technology. Research for this paper was supported with funding from Qualcomm Incorporated and summer research grants from UC-Berkeley Law School and George Washington University Law School.
[**] John Marshall Harlan Research Professor of Law, George Washington University Law School.

ABSTRACT

US and EU privacy law diverge greatly. At the foundational level, they diverge in their underlying philosophy: In the US, privacy law focuses on redressing consumer harm and balancing privacy with efficient commercial transactions. In the EU, privacy is hailed as a fundamental right that trumps other interests. Even at the threshold level – determining what information is covered by the regulation – the US and EU differ significantly. The existence of personal information -- commonly referred to as "personally identifiable information" (PII) – is the trigger for when privacy laws apply. PII is defined quite differently in US and EU privacy law. The US approach involves multiple and inconsistent definitions of PII that are often quite narrow. The EU approach defines PII to encompass all information identifiable to a person, a definition that can be quite broad and vague. This divergence is so basic that it significantly impedes international data flow. A way to bridge the divergence remains elusive, and many commentators have generally viewed the differences between US and EU privacy law as impossible to reconcile.

In this essay, we argue that there is a way to bridge these differences at least with PII. We contend that a tiered approach to the concept of PII (which we call "PII 2.0") represents a superior way of defining PII than the current approaches in the US and EU. We also argue that PII 2.0 is consistent with the different underlying philosophies of the US and EU privacy law regimes. Under PII 2.0, all of the Fair Information Practices (FIPs) should apply when data refers to an identified person or where these is a significant risk of the data being identified. Only some of the FIPs should apply when data is merely identifiable, and no FIPs should apply when there is a minimal risk that the data is identifiable. We demonstrate how PII 2.0 advances the goals of both US and EU privacy law and is consistent with their different underlying philosophies. PII 2.0 thus begins the process of bridging the current gap between US and EU privacy law.

EXECUTIVE SUMMARY

"Personal information" is a central concept in privacy regulation around the world. Personal data -- commonly referred to as "personally identifiable information" (PII) -- is foundational to any privacy regulatory regime because it serves as a jurisdictional trigger. If there is PII, privacy laws apply. If PII is absent, privacy laws do not apply. The concept of PII plays this role in privacy law in the United States (US) and European Union (EU) alike.

A major problem with PII is that it is defined so differently in US and EU privacy law. This divergence accounts for one of the major incongruities between the US and EU privacy regulatory regimes, and it impedes international data flow. More generally, there remains a seemingly intractable philosophical difference between US and EU privacy law. In the US, privacy law focuses on redressing consumer harm and balancing privacy with efficient commercial transactions. In the EU, privacy is hailed as a fundamental right that trumps other interests.

In the US, the law provides multiple definitions of PII. In contrast, in the EU, there is a single definition of personal data to encompass all information *identifiable* to a person. Even if the data alone cannot be linked to a specific individual, if it is reasonably possible to use the data in combination with other information to identify a person, then the data is PII.

In this essay, we argue that both the US and EU approaches to defining PII are flawed. It is time for a new model, which we term "PII 2.0." We contend that a new tiered approach to the concept of PII can bridge the differences between the US and EU approaches. PII 2.0 is also consistent with the different underlying philosophies of the US and EU privacy law regimes.

PII 2.0

PII 2.0 places information on a continuum that begins with no risk of identification at one end and ends with identified individuals at the other. It divides this spectrum into three categories, each with its own regulatory regime:

> *Identified Data.* Information refers to an identified person when it singles out a specific individual from others. Information that brings a substantial risk of identification of an individual should be treated as referring to an identified person.
>
> *Identifiable Data*. Information in the middle of the risk continuum relates to an *identifiable* individual when there is some non-remote possibility of future identification. The risk level for such information is low to moderate.
>
> *Non-identifiable Data*. At the other end of the risk continuum, *non-identifiable* information carries only a remote risk of identification. Such data cannot be said to be relatable to a person, taking account of the means reasonably likely to be used for identification.

PII 2.0 conceives of identifiability as a continuum of risk rather than as a simple dichotomy. A clear way to demonstrate the functioning of this new approach is by considering the applicability of FIPs. The basic toolkit of FIPs includes the following: (1) limits on information use; (2) limits on data collection, also termed data minimization; (3) limits on disclosure of personal information; (4) collection and use only of information that is accurate, relevant, and up-to-date (data quality principle); (5) notice, access, and correction rights for the individual; (6) the creation of processing systems that the concerned individual can know about and understand (transparent processing systems); and (7) security for personal data.

When information refers to an *identified* person, all of the FIPs generally should apply. For the category of *identifiable*, it is not appropriate to treat such information as fully equivalent to identified data. Nonetheless, some protections are in order because there is a risk of linkage to a specific individual. Although all FIPs should not apply to identifiable data, there are three that are applicable to identifiable data: security, transparency, and data quality. For the category of *non-identifiable* data, the information would not be protected by the FIPs

PII 2.0 and EU Privacy Law

PII 2.0 can bridge the chasm between US and EU privacy law despite the fundamental divergence in the underlying philosophy of these two regulatory regimes. PII 2.0 fits with the US harm-based approach because its tiered approach provides more privacy protection when there is a greater risk of harm. This approach contrasts with the EU approach to PII which defines it quite broadly and provides a full suite of rigorous protections to a wide array of data even when there is little risk of harm.

How, then, can PII 2.0 be consistent with the philosophy of EU privacy law? PII 2.0 enhances the protection of privacy by creating an incentive for companies to keep information in the least-identifiable form. Moreover, several of the FIPs such as notice and access rights effectively force companies to keep data in identifiable form in order to administer these rights. Ironically, with many forms of identifiable data, protecting them with the full set of FIPs can undermine privacy protection by making the data less secure and more likely to be processed in more ways. Therefore, for the goal of protecting privacy, having data kept in identifiable rather than identified form is a significant plus, and PII 2.0 encourages keeping data in this format. PII 2.0 would thus strengthen privacy protection in the EU.

PII 2.0 also builds on evidence that EU is evolving past a simple one-size fits-all privacy protection regime. For example, Article 10 of the Proposed Data Protection Regulation provides that data controllers need not collect more personal information to identify the data subject for the mere purpose of complying with the proposed regulation. Moreover, in the context of its opinion on geolocational data, the Article 29 Working Party has taken some initial steps on the road to PII 2.0. It has called for flexibility in the protections needed for data about WiFi routers as opposed to real-time tracking of a mobile device itself. Thus, having different categories of data protected differently is consistent with the EU approach, especially when the categories enhance the overall protection of privacy. PII 2.0 can thus serve as a bridge in the current gap between US and EU privacy law.

I. Introduction

"Personal data" is a central concept in privacy regulation around the world. It defines the scope and boundaries of many privacy statutes and regulations. Numerous federal and state statutes in the United States turn on the definition of "personal data."[1] Personal data is commonly referred to as "personally identifiable information" (PII), and we will thus use the terms interchangeably.

PII is foundational to any privacy regulatory regime because it serves as a jurisdictional trigger: If there is PII, the laws apply. If it is absent, there is no applicability of the privacy regulation in question. The concept of PII also plays a similar role in European Union (EU) privacy law. These laws share the same fundamental assumption—that in the absence of PII, there are no privacy rights. For this reason, privacy regulation focuses on the collection, use, and disclosure of PII, and leaves non-PII unregulated.

Given PII's importance, it is surprising that it lacks a uniform definition. In the US, the law provides multiple explanations of this term. In our previous work, we demonstrated the shortcomings of all of these concepts.[2] In the US, moreover, a commonality in the approaches to PII is to focus on whether the data is actually linked to an identified person. In contrast, in the EU, there is a single definition, and one that defines PII broadly to encompass all information that is *identifiable* to a person. Even if the data alone cannot be linked to a specific individual, if it is reasonably possible to use the data in combination with other information to identify a person, then the information is PII.

The fact that PII diverges so much between US and EU privacy law poses significant difficulties for international data transfers between the US and EU. These exchanges are highly significant, in turn, for global commerce. The European Union is the most important bilateral trade area for the United States. Indeed, the economic relationship between the US and the EU is the largest in the world.[3] According to one estimate from the European Commission, moreover, over half of the EU-US cross-border trade in services depends on the Internet."[4] As a consequence, barriers to "information communication services" will have an impact not only on that sector itself, but other business sectors involved in bilateral EU-US cross border trade.[5]

A large amount of data is merely *identifiable*, but the people to whom the data pertains are not currently *identified*. In a highly significant swath of US privacy law, this information falls outside privacy regulation. In the EU, however, this data is fully regulated pursuant to the rigorous protections of the EU Data Protection Directive. The same result follows under the

[1] Examples of federal laws include the Children's Online Privacy Protection Act, the Gramm-Leach Bliley Act, the HITECH Act, and the Video Privacy Protection Act. Examples of state laws include California's Song-Beverly Credit Card Act and the numerous state breach notification laws.

[2] 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).

[3] William H. Cooper, EU-U.S. Economic Ties: Framework, Scope and Magnitude, Congressional Research Service (Apr. 2, 2013).

[4] European Commission, Commission Staff Working Document, Impact Assessment Report on the future of EU-US trade relations, *Accompanying the document* Recommendation for a Council Decision authorising the opening of negotiations on a comprehensive trade and investment agreement, called the Transatlantic Trade and Investment Partnership, between the European Union and the United States of America, {COM(2013) 136 final},{SWD(2013) 69 final}, n.11, 8 (March 12, 2013).

[5] Id.

more recent EU Proposed General Data Protection Regulation of 2012. This fundamental incongruity in the US and EU regulatory regimes creates significant confusion and impediments to information flow and use.

In previous work, we focused on the approach to PII in US privacy law and criticized the law's disjointed, inconsistent, and often overly narrow definitions of PII. To make privacy law effective for the future, we developed a new conception, PII 2.0, which avoids the problems and pitfalls of current approaches. The key to our model is to build two categories of PII, "identified" and "identifiable" data, and to treat them differently.[6] This approach permits tailored legal protections built around different levels of risk to individuals.

In this Article, we argue that PII 2.0 can do more than serve as the most workable approach for US privacy law. It also would function well for EU privacy law and help harmonize the significantly divergent approaches between US and EU privacy law. This conclusion may appear surprising; it is also far from apparent from our previous work.

EU privacy law functions differently than US privacy law and has different underlying goals and structure. Among many differences, EU law views privacy as a fundamental right, and US law considers it one interest that is balanced against others. It may even be secondary to other concerns, such as freedom of speech. In the EU, privacy law is viewed in broad terms and expressed in omnibus laws that regulate the public and private sectors alike. In the US, privacy law is regulated through narrow sectoral laws that focus on specific industries or specific contexts for the use of personal data. And in the EU, privacy law forbids personal data processing in the absence of a legal basis. In the US, in contrast, the general approach is to allow personal data processing unless it causes a legal harm or is otherwise restricted by law. Given these differences, it is no surprise that EU privacy law has a much broader definition of PII than US privacy law.

Attempts to harmonize US with EU privacy law by turning EU privacy law into a US-style approach, or vice versa, will prove to be non-starters. Both the US and EU are deeply committed to their respective approaches. While policymakers and scholars have been trying for nearly two decades to bring US and EU privacy law closer together, the new EU Proposed Data Protection Regulation could push the US and EU even further apart.

In our view, PII 2.0 can serve as a foundational step in overcoming the differences between US and EU privacy law. In this Article, we set forth the argument for how a tiered approach to the concept of PII can advance the underlying philosophy and goals of EU privacy law and serve as a bridge over trans-Atlantic privacy differences. We will demonstrate how PII 2.0 is a way forward toward harmonization.

II. DEFINING PII ON BOTH SIDES OF THE ATLANTIC

Understanding the modern landscape of information privacy law requires a comparative focus. Legal forces outside the US have significantly shaped the governance of information privacy. In particular, the EU has played a major role in international decisions involving this area of law. This role has been bolstered by its laws granting EU member states the authority to

[6] Schwartz & Solove, supra note 2, at 1877–83. As we will discuss below, as part of PII 2.0's harmonization effort, we leave unchanged the EU category of "sensitive" data.

block transfers to third party nations, including the US.

A. THE EU: FROM THE DIRECTIVE TO THE PROPOSED REGULATION

In the EU, the current framework for defining personal information includes both the Data Protection Directive, which has been in effect since 1995, and a Proposed Regulation, which was released in 2012, and which EU institutions are now debating. The Data Protection Directive, which is currently in force, sets out common rules for data protection in EU member states and requires these countries to enact legislation that follows its standards. Although the Directive employs the term "personal data," this term serves the same function as PII, and in this article, we treat the legal terms "personal data" and "PII" as functional equivalents.

Under both the Directive and Proposed Regulation, the EU approach takes a broad approach to defining PII. The definition turns on whether a natural person is capable, whether directly or indirectly, of identification through a linkage, or some other reference to available data. In the EU, information that is identified or identifiable receives the same level of legal protection.

1. The EU Data Protection Directive

The EU Data Protection Directive uses the term "personal data" and defines it as "information relating to an identified or identifiable natural person."[7] The Directive does not specifically define "identified." Under an EU Directive, the law of member states then becomes determinative. It is generally established that a person falls in the "identified" category if a party can use information relating to her to determine her specific identity**.** In analyzing the term under German law, Ulrich Dammann states, "A person is identified when it is clear that the data relate to the person and not to another."[8] As Rosemary Jay writes in her treatise on UK privacy law, "A person becomes identified where there is sufficient information either to contact him or to recognise him by picking him out in some way from others and know who he/she is."[9]

The EU Data Protection Directive is more specific regarding its definition of "identifiable." It explains that an "identifiable" person is "one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."[10] As additional definitional assistance, the Directive in its Recital 26 explains that in determining whether a person is identifiable, "account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person."[11] This approach follows a long-standing paradigm in German federal data protection law.

Both identified and identifiable information fall squarely within the scope of EU data privacy law, and they are treated in the same fashion. The definition functions as an "on"

[7] Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, art. 2(a), 1995 O.J. (C 93) ("Data Protection Directive").

[8] Ulrich Dammann, § 3 Weitere Begriffsbestimmungen, in Bundesdatenschutzgesetz: Kommentar 310 (Spiros Simitis, ed. 7th ed. 2011).

[9] Rosemary Jay, Data Protection Law and Practice 172 (4th ed. 2012).

[10] Data Protection Directive, supra note 7, at art. 2(a).

[11] Id. at recital 26.

switch for the application of EU data protection law. The Directive specifies obligations on the "data controller," rights for the "data subject," and robust protections for the personal data. As a matter of terminology, EU data privacy law refers to the entity that collects and uses personal data as the "data controller" and the individual whose data is involved as the "data subject."

Once information falls under this regime, a full suite of obligations, rights, and protections is triggered. The general default rule is that the collection and processing of personal data must be for "specified, explicit and legitimate purposes."[12] These purposes may be ones to which the personal data subject has consented, necessary to protect the data subject's vital interests, in the public interest, or in the legitimate interests of the data controller unless they interfere with a data subject's fundamental right to privacy.[13]

The Directive also provides data subjects with a right to control the use of their personal data. Data subjects are to be informed about the entities that collect their personal information, how it will be used, and to whom it will be transferred.[14] Their interests under the Directive also permit data subjects to access their personal data and to correct inaccurate information in their records.[15] The Directive requires that data subjects must provide affirmative consent before their personal data is processed, used, or disclosed. Consent must be unambiguous and freely given. Data subjects have a right to object to a data controller's use of personal data.[16] Data subjects also have a right not to be subject to certain decisions made solely based on the automated processing of data.[17] They are to be informed of the logic used in the automatic processing of that data.

As for data controllers, the Directive imposes a number of obligations on them beyond those that follow from the rights of data subjects. To begin with, data controllers may not process personal information collected for one purpose in ways incompatible with that purpose.[18] Data must be kept accurate and up to date.[19] Data controllers cannot keep personal information for longer than necessary for the purposes for which it was collected.[20] Data must be kept secure.[21]

In additions to these rights and obligations, the Directive also mandates additional protections for certain categories of personal data, or before certain actions may be taken with personal data. There are special restrictions on the use of "sensitive data," which includes data about "racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union memberships, and the processing of data concerning health or sex life."[22] Data controllers must notify supervisory authorities before engaging in many kinds of data processing.[23] The Directive also restricts the transfer of personal data to other countries. Personal data may be transferred only to countries with an "adequate level of protection" of privacy.[24]

[12] Id. at art. 6(b).
[13] Id. at art. 7.
[14] Id. at art. 10.
[15] Id.
[16] Id. at art. 14.
[17] Id. at art. 15.
[18] Id. at art. 6(b).
[19] Id. at art. 6(d).
[20] Id. at art. 6(e).
[21] Id. at art. 17.
[22] Id. at art. 8.
[23] Id. at art. 18
[24] Id. at art. 25.

In sum, in the EU, information that is identified or identifiable falls under the definition of "personal data." The consequence is to trigger a wide range obligations, rights, and protections. As we will now see, the EU Proposed Data Protection Regulation also treats identified and identifiable as equivalents. Its new term of art, however, is "indirectly identified" rather than the Directive's term, "identifiable."

2. The Proposed Regulation

The EU is now in the process of replacing the Directive with a Data Protection Regulation. In January 2012, the EU Commission released a draft version of this document, its Proposed General Data Protection Regulation.[25] This development marks an important policy shift. In EU law, while a directive requires Member States to pass harmonizing legislation, a regulation establishes directly enforceable standards. As Christopher Kuner explains, "a regulation leads to a greater degree of harmonization, since it immediately becomes part of a national legal system, without the need for adoption of separate national legislation; has legal effect independent of national law; and overrides contrary national laws."[26] Due to its directly binding effect, the Data Protection Regulation, if approved, will be even more important-- and from its first day of enactment-- than the Directive.

The Proposed Data Protection Regulation generally builds on the approach of the Directive, but contains some notable changes. Instead of a concept of "identifiable," it relies solely on a concept regarding a person being "indirectly identified."[27] The Proposed Regulation also provides additional examples of the kinds of linkages that tie information, whether directly or indirectly, to a "data subject." The new examples refer to "location data," "an online identifier," and "genetic" identity. Their impact is to modernize and expand the categories from the 1995 Directive.[28] This expansion of the categories to indirect identification demonstrates the EU philosophy of taking a broad approach to defining personal information. Most crucially,

[25] Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data (General Data Protection Regulation) ("Proposed Regulation") (Jan. 25, 2012).

[26] Christopher Kuner, *The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law*, 11 PRIVACY & SECURITY L. REP. 1, 3 (2012).

[27] Proposed Regulation, supra note 25, at art. 4(1).

[28] The Proposed Regulation's key language comes in its definition of "data subject." This term refers to the individual whose personal data is processed and who can be identified. The relevant language at Article 4 is worth citing. A data subject is:

> an identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.

Id. As this language indicates, the Proposed Regulation drops the Directive's language about "identifiable," but retains its idea of indirect identification. It also expands the Directive's categories of the kinds of information that can lead to identification of a data subject. These now include an "identification number," "location data," and "online identifier." In a similar fashion, the Proposed Regulation expands the Directive's list of specific items that can identify by adding information pertaining to "genetic" identity. This additional factor reflects the ability of modern science to de-identify certain kinds of information. A reference to genetic information, just as the other factors listed in the Proposed Regulation, can lead to the identification of a specific person.

it still uses the concept of "personal data" as the "on" switch for triggering a full suite of obligations, rights, and protections.

At the same time, the Proposed Regulation contains helpful indications of the need for flexibility in deciding when personal information does or does not exist. For example, its Recital 24 provides important limitations on the Proposed Regulation's concept of "indirect identification." Recital 24 initially notes that use of "online services" may lead individuals to "be associated with online identifiers provided by their devices, applications, tools and protocols, such as Internet Protocol addresses or cookie identifiers."[29] This association may lead to the creation of profiles through "information received by the servers" and lead to identification of individuals.

The Recital also observes that these kinds of associations need not create identifiable information. Recital 24 states: "[I]dentification numbers, location data, online identifiers or other specific factors as such need not necessarily be considered as personal data in all circumstances."[30] This language indicates the potential under the Regulation for needed flexibility in determining the presence or absence of personal information.

Recital 24 of the Proposed Regulation adopts a context-specific analysis for deciding whether or not personal data is present. While the Proposed Regulation is specific about the need for flexibility, the ultimate test regarding "identifiability" is the same under both the Directive and Proposed Regulation. Both documents use the same language, and we have already cited it above: the analysis must consider "all the means likely reasonably to be used either by the controller or by any other person to identify" the individual.[31] The repetition of the language in both documents indicates a certain continuity of approach in the EU to defining personal data. The necessary analysis must consider the likelihood that certain steps, such as combining bits of scattered data or re-identifying non-personal information, will actually be taken.

Thus, an identified person in the EU is one that can be singled out, whether directly or indirectly, through a linkage to information that references her or him. In a fashion that is consistent with the Directive's approach, the Proposed Regulation offers an expansionist approach to defining personal information. The critical analysis in the EU remains focused on whether a natural person is capable of identification, based on an analysis of all means likely to be used and by reference to available data. Finally, Recital 24 of the Proposed Regulation points to possible flexibility in deciding when personal information is and is not present.

The breadth of the EU approach has both benefits and drawbacks. The primary benefit is that hardly anything escapes EU privacy regulation. There are few gaps and inconsistencies under the EU approach, a stark contrast to the US where such gaps and inconsistencies are legion. But there is also a primary drawback to the EU approach. Under both the Directive and Proposed Regulation, information is treated as the same whether it refers to an identified individual, or one who can be "indirectly identified," that is, an identifiable person. The difficulty is that there is a broad continuum of identifiable information that includes different types of anonymous or pseudonymous information. Different levels of effort are required to identify information, and various risks are associated with the possible identification of data. To place all such data into the same conceptual category as data that currently relate to an

[29] Id. at recital 24.

[30] Id.

[31] Id. at recital 23; Data Protection Directive, supra note 7, at recital 26.

identified person is an approach that lacks nuance.

B. THE US: A LACK OF A UNIFORM STANDARD

Instead of defining personal information in a coherent and consistent manner, privacy law in the US offers multiple competing definitions. As an initial matter, the law in the US at times uses "identifiable" as synonymous with "personal data." For example, common usage sometimes sees information privacy law in the US as regulating "personally identifiable information." The Google Ngram database shows that "personally identifiable information" has even become a highly popular way to express the concept of "personal information" since 1995.[32] When information is identifiable enough to fall within the scope of a particular statute relies, however, on the specific definition within each privacy statute.

In American law, there are three predominant approaches to defining personal information. These are (1) the "tautological" approach, (2) the "non-public" approach, and (3) the "specific-types" approach. The tautological approach is an example of a standard, or an open-ended decision-making tool. The classic example of a standard would be an instruction to "drive at a reasonable speed," or, in the law of negligence, to take the precautions of a reasonable person. Under the tautological approach, US privacy law simply defines "personal" as meaning any information that identifies a person. The Video Privacy Protection Act of 1988 (VPPA) neatly demonstrates this model.[33] The VPPA, which safeguards the privacy of video sales and rentals, defines "personally identifiable information" as "information which identifies a person."[34] For purposes of the statute, information that identifies a person becomes "personal identifiable information" and falls under the statute's jurisdiction once linked to the purchase, request, or obtaining of video material.

A second approach is to focus on non-public information. The non-public approach seeks to define personal information by focusing on what it is *not* rather than on what it is. Instead of saying that personal information is simply that which identifies a person, the non-public approach separates out information that is publicly accessible or information that is purely statistical. Information that falls in these categories is not personal information, but the relevant legislation often does not explore or develop the logic behind this approach.

The Gramm-Leach-Bliley Act of 1999 (GLBA) epitomizes this approach by defining "personally identifiable financial information" as "nonpublic personal information."[35] The statute fails to define "nonpublic," but presumably this term means information not found within the public domain. The FTC Regulations to the GLBA explain this term in more detail, but leave confusion as to the extent to which data cannot be "public," that is accessible to the public, for it to be classified as "non-public" in the sense of the statute. The applicable regulations also sweep in "any information" that a consumer provides on a financial application, which seems to relax the concept of "nonpublic."

In an illustration of another aspect of the "nonpublic" approach, the Cable Communications

[32] *Google Books Ngram Viewer*, http://books.google.com/ngrams (last visited Mar. 13, 2013).

[33] 18 U.S.C. § 2710 (2006).

[34] Id. § 2710(a)(3). The VPPA prohibits "videotape service providers" from knowingly disclosing personal information, such as the titles of items rented or purchased, without the individual's written consent. It defines "videotape service providers" in a technological neutral fashion to permit the law to be extended to DVDs. Id. § 2710(a)(4).

[35] 15 U.S.C. § 6809(4)(A) (2006).

Policy Act of 1984 defines PII as something other than "aggregate data."[36] This statute, which protects the privacy of subscribers to cable services, views PII as excluding "any record of aggregate data which does not identify particular persons."[37] By aggregate data, the Cable Act presumably means purely statistical information that does not identify specific individuals.[38]

The third approach of American privacy law is to list specific types of data that constitute personal information. In the context of the specific-types approach, if information falls into an enumerated category, it becomes per se personal information under the statute. State data breach notification laws take this approach. These statutes, which forty-six states have now enacted, require a business to notify affected individuals should an information security breach occur. The typical trigger for these laws is the unauthorized acquisition or access to unencrypted personal information. These laws then typically define personal information through the specific-types approach. As an illustration, the Massachusetts breach notification statute requires that individuals be notified if a specific set of their personal information is lost or leaked.[39] The Massachusetts act defines personal information as a person's first name and last name, or first initial and last name in combination with a social security number, driver's license number, financial account number, or credit or debit card number.[40] As Lisa Sotto notes in her privacy law treatise, "Many states have varied the definition of personal information" to include not only the elements listed above, as found in the Massachusetts law, but also "any number of other potentially identifiable elements."[41]

Under all these approaches, however, the presence of personal information is a prerequisite to trigger data breach notification. Even more specifically, and as Sotto writes, these laws typically require the presence of "a state's resident's first name, or first initial and last name" in combination with the other elements.[42] Only a few states have a trigger in their data security breach notification laws other than first name or initial and last name. As examples of this minority approach, Georgia, Maine, and Oregon have general "savings clauses" that extend protection to data elements even when they are not connected to a person's name if the information would be sufficient to permit identity theft.[43] These states are leading the way for better, next-generation data breach notification laws.[44] In the meantime, however, the vast majority of these laws focus only on identified individuals. The flaw of these statutes is apparent: certain information, beyond first name, or first initial, and last name is readily capable of identifying a specific individual. A breach of such information should be covered by these statutes.

One can also point to a broader flaw in the approach of US law to PII. Although there are a few exceptions, as a general rule, PII in the US is largely limited to instances where data refers to an identified individual. There is a clear contrast with PII in the EU, where data protection law extends expansively to any data that is identifiable – i.e., that could possibly be linked to an

[36] 47 U.S.C. § 551(a)(2)(A) (2006).

[37] Id.

[38] The number of Comcast customers in Virginia who subscribe to HBO is an example of aggregate data under the Cable Act.

[39] Mass. Gen. Laws Ann. ch. 93H (West 2007).

[40] Id. § 1.

[41] LISA J. SOTTO, PRIVACY AND DATA SECURITY LAW DESKBOOK 15-4 (2011).

[42] Id. at 15-5.

[43] Ga. Code Ann. § 10-1-911(6); Me. Rev. Stat. Ann. tit. 10, §1347(6); Or. Rev. Stat. § 646A.602(11).

[44] For a chart exploring and categorizing the various PII definitions in different state data security breach notification laws, see DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 176–78 (2013).

individual.

C. Personal Information: A Problem on Both Sides of the Atlantic

The approaches to defining PII in the US and in the EU are all flawed in significant ways. Stated succinctly, we find the EU approach to be too expansionist, and the US approach too reductionist -- and both of these approaches have problematic consequences. Moreover, the divergence between these approaches raises severe problems in a global economy which depends upon the free flow of data between the US and EU.

If PII or personal data were a small dimension of privacy regulation, such problems might be isolated and worked around. But the definition of PII is a foundational issue that implicates the scope of privacy regulation. Before even considering differences in *how* data is protected in the US and EU, we must address differences in *what* data is protected under these two privacy regimes. The state of disarray for this issue points to the critical importance of revising the definition of PII in the US and EU alike.

1. Evaluating the EU Approach

The benefit of the EU approach to personal information is that it recognizes the expanding ability of technology to re-identify information and to link scattered crumbs of information to a specific individual. The notion of taking "account … of all the means likely reasonably to be used" to identify a person points to a flexible, context-based analysis.[45] As noted, moreover, the Proposed Regulation states that "identification numbers, location data, online identifiers or other specific factors as such need not necessarily be considered as personal data in all circumstances."[46] Here, too, there is an indication that an evaluation should consider whether possible steps in combination or re-identification of data are likely to be taken.

Despite this promise, the EU's definition of personal information risks sweeping too broadly. Much depends on judgments about open-ended factors such as "the means likely reasonably to be used." In its opinion "On the Concept of Personal Data," the Article 29 Working Party develops a liberal approach to when information falls within the scope of this term as well as some problematic illustrations of its chosen principles.[47] The Article 29 Working Party is an important group of EU national data protection commissioners. Under the Directive, it has an advisory role in contributing to "the uniform application" of national privacy law.[48] As such, its opinions on issues such as the definition of personal data provide a window into EU privacy law.

In its 2007 opinion, the Article 29 Working Party presents a number of general overarching principles for deciding when personal information is present. It begins with the idea that the Directive contains a "broad notion of personal data."[49] This document's wide notion in turn furthers the Directive's objective, which is to protect "the fundamental rights and freedoms" of individuals. As a consequence, one should not "unduly restrict the interpretation of the

[45] See supra note 31.
[46] Proposed Regulation, supra note 25, at recital 24.
[47] Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, WP 136 (June 20, 2007).
[48] Data Protection Directive, supra note 7, at art. 30(1)(a).
[49] Article 29 Data Protection Working Party, supra note 47, at 4.

definition of personal data."[50] Moreover, the core insight of the Article 29 Working Party is sound: in looking at whether information is personal data, the analysis must be a "dynamic one" that considers, among other factors, the "state of the art" of the relevant technology and how it is likely to advance over the information's life cycle.[51]

These principles form elements of a useful approach to assessing the presence or absence of "personal data." Yet, the Article 29 Working Party's own interpretation of these concepts is far from unproblematic. First, the Working Party redefines the crucial activity in question to decisionmaking based on specific characteristics of the person.[52] This is a different issue than that of identifiability. Second, the Working Party views information as *per se* identifiable if the ultimate purpose of the data controller is to identify *some* of the parties in the database.[53] This approach further shifts the analysis away from a risk or a harm analysis, or, indeed, from an estimation of the likelihood of identification. Its focus is on the moment of collection and data and the processing purpose.

The Article 29 Working Party's analysis of web tracking illustrates the first point. The Working Party concludes that a unique identifier assigned to a computer on the Web creates personal information because "web traffic surveillance tools make it easy to identify the behaviour of a machine and, behind the machine, that of its user."[54] The Working Party states:

> Without even enquiring about the name and address of the individual it is possible to categorise this person on the basis of socio-economic, psychological, philosophical or other criteria and attribute certain decisions to him or her since the individual's contact point (a computer) no longer necessarily requires the disclosure of his or her identity in the narrow sense.[55]

This analysis moves information into the "identifiable" category because of a link of an identifier to a computer and decisionmaking about characteristics of a person ("socio-economic, psychological or other criteria").

This focus seems related to a particular concern in the EU about decisions about a person based on automatic means.[56] As noted above, the Directive provides protection against decisionmaking based solely on automated decision-making. The Proposed Regulation follows this path; it requires limits on automated decisionmaking and ties its concern to contemporary concerns about profiling. Its Article 20 states:

> Every natural person shall have the right not to be subject to a measure . . . which is based solely on automated processing intended to evaluate certain personal aspects relating to this natural person or to analyse or predict in particular the natural person's performance at work, economic situation, location,

[50] Id. at 5.
[51] Id. at 15.
[52] See id. at 12.
[53] See id. at 13.
[54] Id. at 14.
[55] Id.
[56] Id. at 5 ("The processing of personal data by non-automatic means is only included within the scope of the Directive where the data form part of a filing system or are intended to form part of such system (Article 3).").

> health, personal preferences, reliability or behaviour.[57]

In our view, however, the necessary analysis should be different than that of the Article 29 Working Party.

Web tracking through the placing of alphanumerical codes on an individual's computer raises a host of complex issues and, in some cases, significant risk of privacy violations. For example, while contemporary advertising networks may not know people's names, identification of individuals is nonetheless possible in many cases. In certain circumstances, enough pieces of information can be associated with a person through her use of a computer to make the process of identification a genuine possibility. At other times, this identification will not be such a possibility, which means there is not use of personal information.

Privacy harms require data use or disclosure pertaining to a specific individual who is identified or reasonably identifiable. In our view, identified information is present when a person's identity has been ascertained, or when there is substantial risk of identification of a specific person by a party who is likely to obtain that information. Targeted marketing that categorizes persons on the basis of socio-economic and other criteria can raise issues about consumer protection and discrimination towards certain groups. Yet, unless this gathering of information creates data that is reasonably capable of being linked to her identity, it does not create identified information. Depending on the precise safeguards that the web tracking company takes, this information may only be identifiable, or even non-personal data.

As a second problematic element in this 2007 opinion, the Article 29 Working Party views any information that is stored as *per se* identifiable if the ultimate purpose of the data controller is to identify *some* of the parties in the database.[58] The opinion's specific examples concern video surveillance, dynamic IP addresses, and the recording of graffiti tags by a transportation company.[59] The problem is that the Working Party's approach confuses collection and stated purpose with identifiability. As a result, it sees identifiable information as present even in circumstances when most or even all of the information in question is never identified.[60]

The Working Party's logic is straightforward. It frontloads the analysis in a fashion that turns the *collection* of information and the overall stated *purpose* into the decisive event for analysis of whether personal data are present. It argues that if "the purpose of the processing implies the identification of individuals, it can be assumed that the controller or any other person involved have or will have the means 'likely reasonably to be used' to identify the data subject."[61] In each case, so long as the ultimate intention is to link some of these data to individuals, all of the information, including those that are never tied to any person, are treated as personal data.

For the Working Party, none of the information collected may be identified. Nonetheless, its model turns all of the information, from the moment of its collection, into identifiable data, that, moreover, in the EU generally receives the same status as identified information. This result is further illustrated by a final example in the Article 29 Working Party's opinion on personal data. In it, the Working Party considers "key-coded data," which is typically used in

[57] Proposed Regulation, supra note 25, at art. 20(1).

[58] Article 29 Data Protection Working Party, supra note 47, at 16.

[59] Id. at 16–17.

[60] Id. at 16.

[61] Id.

clinical trials with medicines.[62] In such clinical trials, a key permits identification of individual patients. Such identification is needed if, for example, medicines turn out to be dangerous and participants in a clinical trial must receive treatment as a consequence. The Article 29 Working Party views identification as "something that *must* happen under certain circumstances."[63] It therefore concludes, "In this case, one can conclude that such key-coded data constitutes information relating to identifiable natural persons for all parties that might be involved in the possible identification and should be subject to the rules of data protection legislation."[64] Yet, some or even all of this data might *never* be identified, and in those cases the party who has access to the data but not the keys is handling information that for it is functionally non-personal information. A data controller might maintain keys encrypted with strong institutional safeguards to prevent access to the key-coded data unless carefully defined events occur. In certain circumstances, therefore, the possibility of identification may be highly remote for the party who has access only to key-coded data.

A final example will demonstrate how the EU's concept of personal data skimps on analysis of whether data is reasonably likely to be identified. This illustration is drawn from Christopher Kuner's treatise on European Data Protection Law. Kuner finds that the EU concept of "identifiability" means that a set of data can be matched to a particular person by some party somewhere regardless of whether the data controller can do so. His example concerns "all males over 50 living in city X who are physicians, have two daughters, listen to Verdi operas and have vacation houses in the South of France."[65] Such information is personal data "even if the data controller could not, with reasonable effort, create a link to an identifiable individual, as long as any party could do so."[66]

Kuner's discussion indicates something analytically troubling about EU law. Assume that the data controller in question cannot, as Kuner posits, create a link to an individual "with reasonable effort." As a further condition, assume that the data controller also institutes strong measures to keep this data secure and promises never to share it with any parties who can link it to an individual. It seems odd under these conditions to grant this information the full set of privacy protections as identified data.

2. Evaluating the US Approach

There are also considerable flaws in the US approach to personal information. Recall that there is no single US definition of this term, but three approaches: the tautological approach, the non-public approach, and the specific-types approach. To begin with, there is a distinct flaw in having three available classifications. As a consequence of these multiple possibilities, the same information may or may not be personal data under different statutes and in different processing contexts. At a minimum, the presence of three definitions increases the regulatory maze and associated compliance costs for regulated entities.

Moreover, each of the three approaches in the US has its own flaws. The tautological approach fails to define personal information or explain how it is to be singled out. At its core, this approach simply states that personal information is personal information. As a result, this definition is unhelpful in distinguishing personal data from non-personal data.

[62] Id. at 19.

[63] Id. at 20.

[64] Id.

[65] CHRISTOPHER KUNER, EUROPEAN DATA PROTECTION LAW 92 (2d ed. 2007).

[66] Id. at 93.

The initial problem with the non-public approach is that it does not map onto whether the information is in fact identifiable. The public or private status of data often does not match up to whether it can identify a person or not. For example, a person's name and address, which clearly identify an individual, nevertheless might be considered public information, as such information is typically listed in telephone books. In many cases, however, individuals have non-public data that they do not want matched to this allegedly public information. Yet, an approach that only protects non-public information as PII might not preclude such combinations.

The second problem with the non-public approach is that it confusingly suggests that if information is public somewhere the parties who process it are not handling regulated data. This confusion arises under the Gramm-Leach-Bliley Act.[67] As we have seen, this statute regulates "non-public personal information" (NPPI). NPPI is defined as "personally identifiable financial information" that a consumer supplies or that is obtained in connection with a transaction involving a financial product or services.[68] In turn, "publicly available information" is defined as "any information that a financial institution has a reasonable basis to believe is lawfully made available to the general public."[69] Sources of this information can include federal, state, or local government records, or widely distributed media.

This approach is more than a little perplexing. The concept of "non-public personal information" may mistakenly suggest to entities that certain information does not fall under the Gramm-Leach-Bliley Act, or other federal privacy regulations, because it is available somewhere from some entity. For example, the *N.Y. Times* has recently reported on individuals using financial information online, including credit scores, as part of the assessment of potential dates and romantic partners.[70] Flaws in the regulation of databrokers and inadequate enforcement mechanisms in the applicable statutes have made a host of financial data widely accessible. Yet, these data should not be considered as NPPI under information privacy statutes, such as the Gramm-Leach-Bliley Act. The information, while perhaps widely available to anyone with access to the Internet and a search engine, should not be seen as available "lawfully," which is a requirement in the FTC's Regulations.

As a potential source of additional confusion, the law may impose obligations on organizations even for *public* information. As an example, the FTC Safeguards Rule notes that *all* customer information must be properly safeguarded once it is in the possession of an entity that falls under GLBA.[71] The Rule requires that these entities protect "the security and confidentiality of customer records and information . . . which could result in substantial harm or inconvenience to any customer."[72] Yet, the concept of NPPI may mistakenly suggest more limited obligations to the regulated entity.

As for the specific-types approach, an initial problem is that it can be quite restrictive in its definition of personal information. The Massachusetts data breach statute defines personal

[67] 15 U.S.C.A. § 6801 et seq.

[68] Id. § 6809(4)(A).

[69] Federal Trade Comm'n, *The Gramm-Leach-Bliley Act* (2001), *available at* http://www.ftc.gov/privacy/glbact/glboutline.htm.

[70] Jessica Silver-Greenberg, *Perfect 10? Never Mind That. Ask Her for Her Credit Score,* N.Y. TIMES (Dec. 25, 2012), http://www.nytimes.com/2012/12/26/business/even-cupid-wants-to-know-your-credit-score.html?pagewanted=all&_r=1&.

[71] 16 C.F.R. § 314.1.

[72]15 U.S.C.A. § 6801(a)-(b).

information to include only a narrow set of data elements: a name plus other elements, such as a Social Security Number, a driver's license number, or a financial account number.[73] This list is under-inclusive: there are numerous other kinds of information that, along with a person's name (or independently), would reveal one's identity. Moreover, most individuals would consider such a data breach to be a significant event and one about which they would wish to be informed.

The Massachusetts version of the specific-types approach also assumes that the types of data that are identifiable to a person are static—- the statute does not cover information that could potentially become personally identifiable. Finally, and as noted above, most data breach statutes have a fixed requirement of a last name and a first name, or the initial of the first name. A leak of many other types of information can reasonably be expected to cause identification of a specific individual. This variant of the specific-types approach is too rigid to adequately protect personal privacy.

As for COPPA, a second example of the specific-types approach, this federal statute has an advantage that data breach notification laws generally lack. COPPA explicitly references FTC rulemaking as a way to expand and adapt its definition of personal information.[74] The FTC has indeed acted to expand the definition of personal information in the statute; its COPPA rule added one element to the statutory concept of personal information, namely, the idea of "a persistent identifier," such as a cookie.[75] However, the FTC's ability to alter the definition of personal information is limited by a requirement that information covered by the statute must permit the "contacting of a specific individual."[76] This language has a tautological element; it front-loads the issue at stake, that is, whether or not the information in question will reference a specific person.

As this analysis shows, the US approach suffers from numerous weaknesses. Overall, it creates inconsistencies and can leave too much information unprotected. As an example, a spokesperson for the online advertising industry has stated that its "tracking doesn't violate anyone's privacy because the data sold doesn't identify people by name."[77] Since there is neither a single model nor shared understanding of "personal information" in the US, the refutation of this claim becomes no simple matter.

The US system is also likely to lead to gaps in protection. For example, whether information can be re-identified depends upon technology and corporate practices that permit the linking of de-identified data with already identified data. As additional pieces of identified data become available, it becomes easier to link them to de-identified data because there are likely to be more data elements in common. The US definitional approach and, in particular, the specific types approach, do not seem well equipped to function in this world of readily available data and context-specific analysis.

3. The Disjunction Between the US and EU Definitions of PII

[73] Mass. Gen. Laws Ann. ch. 93H § 1(a) (West 2007).

[74] 15 USC. § 6501(8)(F) (2006).

[75] 16 C.F.R. § 312.2 (2011).

[76] 15 USC. § 6501(8)(F).

[77] See Julia Angwin & Tom McGinty, *Sites Feed Personal Details to New Tracking Industry*, WALL ST. J. (July 30, 2010), http://online.wsj.com/article/SB10001424052748703977004575393173432219064.html.

The disjunction between US and EU definitions raises problems regarding international transfers of personal data. To understand these difficulties, we should first consider the complex legal structure for judging the permissibility of these transfers under EU law. This analysis requires examination of the current approach under the Data Protection Directive and then the suggested future approach under the Proposed Data Protection Regulation.

In its Article 25, the Directive permits transfers to "third countries," that is, countries outside of the EU, only if these nations have "an adequate level of protection." [78] The EU does not generally consider the US to provide "adequate" privacy protection, and negotiations have taken place to develop mechanisms for US companies to meet the "adequacy" requirement of the Directive's Article 25. The resulting mechanisms include the Safe Harbor, Model Contractual Clauses, and Binding Corporate Rules. The Safe Harbor, which was negotiated between the EU and US Department of State, creates a voluntary self-certification program for US firms. Its mixture of substantive standards combine EU and US privacy requirements, but ends somewhat closer to the EU version of these norms.[79] The EU has approved two sets of Model Contractual Clauses, which simplify the process of crafting data transfer agreements. Finally, Binding Corporate Rules (BCRs) provide another way to meet the Directive's adequacy requirement. Through BCRs, an organization pledges to meet the Directive's standards in its data processing and promise its cooperation with EU data protection authorities.

Article 25 of the Directive calls for an evaluation of adequacy "in the light of all the circumstances surrounding a data transfer operation or set of data transfer operations."[80] Hence, it requires a contextual analysis of the protections in place in the non-EU country. Article 25 of the Directive specifies that "particular consideration shall be given to the nature of the data, the purpose and duration of the proposed processing operation or operations, . . . the rules of law . . . in force in the third country in question and the professional rules and security measures" in that country.[81]

The Proposed Regulation generally continues the approach of the Directive to data transfers outside of the EU. Like the Directive, it allows these transfers when the personal data will receive "an adequate level of protection." It is also more flexible than the Directive in certain regards. The Proposed Regulation notes that an adequacy determination is to be made by the Commission by examination of "the third country, or a territory, or a processing sector within that third country, or the international organisation in question."[82] This language suggests that smaller geographical areas, such as individual states, or particular companies might receive an adequacy determination from the Commission. As a further indication of flexibility, the Proposed Regulation permits transfers if there is a use of "appropriate safeguards … in a legally binding instrument" as part of the transfer, or the use of one of eight possible exceptions.

Under either the Directive or the Proposed Regulation, however, the resulting analysis is complicated by the differences in definitions of PII in the EU and US. Information that is "identifiable" in the EU and, hence, subject to Article 25 of the Directive, or the new provisions for data transfers in the Proposed Regulation, may not receive protection in the US. Such information may not fall within the scope of the applicable definition in the relevant sector. Mechanisms, such as the Safe Harbor agreement, may fail to bridge these differences because of

[78] Data Protection Directive, supra note 7, at art. 25(1).
[79] Paul M. Schwartz, *The EU-U.S. Privacy Collision*, 126 HARV. L. REV. 1980–81 (2013).
[80] Data Protection Directive, supra note 7, at art. 25(2).
[81] Id.
[82] Proposed Regulation, supra note 25, at art. 41.

the lack of agreement about the basic unit of privacy law, which is the concept of "personal data."[83]

Because of the differences in definitions of PII, privacy law in the US or EU has a vastly different jurisdictional scope. These differences persist even in sectors covered by both US and EU privacy law, such as the Safe Harbor agreement. This divergence creates a significant impediment to global commerce and communication.

III. PII 2.0

The existing definitions of personal information, whether in the EU or US, are problematic. Nonetheless, information privacy law should not abandon the concept of PII. If it took this step, privacy law would be left without a means for establishing coherent boundaries on the necessary regulation. Therefore, we reconceptualize the current standards in the EU and US through PII 2.0.

A. AN EXPLANATION OF PII 2.0

Our model places information on a continuum that begins with no risk of identification at one end, and ends with identified individuals at the other. We divide this spectrum into three categories, each with its own regulatory regime: under the PII 2.0 model, information can be about an (1) identified, (2) identifiable, or (3) non-identifiable person. Our three categories divide up this spectrum and provide different regimes of regulation for each. Because these categories do not have hard boundaries, we define them in terms of standards, that is, as open-ended benchmarks rather than as hard-edged "rules."

1. Identified Data

Information refers to an *identified* person when it singles out a specific individual from others. Put differently, a person has been identified when her identity is ascertained. There is general international agreement about the content of this category, albeit not of the implications of being placed in it. For example, in the United States, the Government Accountability Office, Office of Management and Budget, and National Institute of Standards and Technology associate this concept with information that distinguishes or traces a specific individual's identity.[84] In Europe, the Article 29 Working Party states that a person is identified "when, within a group of persons, he or she is 'distinguished' from all other members of the group."[85] This definition also follows that of member states, which, as we have seen, assess whether information relating to a person has determined her specific identity**.** To return to the example in Jay's treatise about U.K. data protection law, a person is identified "where there is sufficient

[83] As a further matter, there has been much criticism of the Safe Harbor Agreement in the EU and beyond. See Lothar Determann, Data Privacy in the Cloud: A Dozen Myths and Facts, 28 Computer & Internet Lawyer 1 (11/Nov. 2011)("The US-EU safe harbor program has been heavily criticized over the years, and the head of a data protection authority in one German state has even called for a rescission of the program."). Despite this criticism, rescission of it seems unlikely at present.

[84] Erika McCallister et al., Nat'l Inst. of Standards and Tech., Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) 2-1 (2010); US Gov't Accountability Office, GAO-08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information 1 n.1 (2010).

[85] Article 29 Data Protection Working Party, supra note 47, at 12.

information either to contact him or to recognise him by picking him out in some way from others and know who he/she is."[86]

EU data protection law also contains special protections for sensitive data. PII 2.0 leaves this special designation in place. The EU Data Protection Directive identifies a special category of data called "sensitive data" that includes data about "racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, . . . health or sex life."[87] This data receives stronger protections than other types of data. The Proposed Regulation also recognizes a similar category. As its Recital 41 states, "Personal data which are, by their nature, particularly sensitive and vulnerable in relation to fundamental rights or privacy, deserve specific protection."[88] Its Article 9 expresses its protections for such "special categories of personal data."[89] Under PII 2.0, identified data that falls into the sensitive category continues to receive these special protections pursuant to EU law.

There are also certain instances where *identifiable* information should be treated like information referring to an *identified* person. Information that brings a substantial risk of identification of an individual should be treated as referring to an identified person. In other words, identifiable data should be shifted to the *identified* category when there is a significant probability that a party will make the linkage or linkages necessary to identify a person.

This essential subcategory requires assessment of the means likely to be used by parties with current or probable access to the information, as well as the additional data upon which they can draw. This test, like those for the other categories, is a contextual one. It should consider factors such as the lifetime for which information is to be stored, the likelihood of future development of relevant technology, and parties' incentives to link identifiable data to a specific person. Here, one of the important factors to be considered are steps that a company takes to keep information from being linked to any specific individual.

2. Identifiable Data

Information in the middle of the risk continuum relates to an *identifiable* individual when specific identification, while possible, is not a significantly probable event. In other words, an individual is identifiable when there is some non-remote possibility of future identification. The risk level for such information is low to moderate. Information of this sort should be regulated differently from an important subcategory of nominally identifiable information, in which linkage to a specific person has not yet been made, but where such a connection is more likely.
As we have discussed above, such nominally identifiable data should be treated the same as identified data.

An example for PII 2.0 of *identifiable* information would be the key-coded medical data that the Article 29 Working Party discusses in its "Opinion on the Concept of Personal Data." Some or all of this information might never be identified. Depending on the risk scenario, there may only a non-remote chance of future linkage to a specific person. As a further example, Kuner's discussion of the Verdi-loving physician may only represent merely identifiable information under PII 2.0. Kuner's hypothetical leaves much open regarding the "data controller." We know only that he himself cannot identify the person to whom the information

[86] Jay, supra note 9, at 172.
[87] Data Protection Directive, supra note 7, at art. 8.
[88] Proposed Regulation, supra note 25, at recital 41.
[89] Id. at art. 9.

relates. If the data controller also has strong measures in place to protect the data from exposure to others, it would not been seen as identifiable, but not identified, under PII 2.0.

In 2012, the FTC Staff Report, *Protecting Consumer Privacy in an Era of Rapid Change*, considered the issue of when information is "reasonably linkable" to a person.[90] In our view, the FTC's approach is helpful regarding this issue of "identifiable" but not "identified." This approach will encourage companies to prevent merely identifiable information from becoming identified information. Citing to our previous work on PII 2.0, the FTC noted that businesses can sometimes re-identify non-PII data and often have incentives to do so.[91] It argued that if companies take three specific steps to minimize linkability, the information should be viewed as non-PII. First, the company must use reasonable means to be confident that the data is de-identified, that is, that it cannot be tied to a specific consumer.[92] Second, the company must publicly commit that it will use the data only in de-identified fashion and not attempt to re-identify it.[93] Finally, if the company makes the de-identified data available to other companies, it must contractually bind the other entities from re-identifying the data and engage in reasonable oversight to monitor compliance with these contracts. [94]

3. Non-identifiable Data

At the other end of the risk continuum, *non-identifiable* information carries only a remote risk of identification. Such data cannot be said to be relatable to a person, taking account of the means reasonably likely to be used for identification. In certain kinds of data sets, for example, the original sample is so large that other information will not enable the identification of individuals.

A simple example of non-identifiable information would be high-level information about the populations of the United States, China, and Japan, and their relative access to telecommunications. At some abstract level, this information refers to persons; it is not merely data about the physical or manmade world, such as the sky is blue, or that Route 95 goes through New Haven, Connecticut. Yet, this information also cannot be linked to a specific person.

Practical methodologies now exist for assessing the risk of identification. In fact, computer scientists have developed metrics for assessing the risk of identifiability of information. For example, Khaled El Emam has identified benchmarks for assessing the likelihood that de-identified information can be linked to a specific person—that is, can be made identifiable.[95] The critical axes in El Emam's work concern the "mitigating controls" available to parties in possession of information, and the likely motives and capacity of outsiders who might seek to tie that information to a person.[96] In addition, computer scientists' ongoing work in developing more secure software offers useful lessons. The relevant need is to focus on: (1) the nature of internal and external threats to a data asset, and (2) the effectiveness of possible

[90] See Federal Trade Comm'n, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Business and Policymakers* iv (Mar. 2012).

[91] Id. at 20.

[92] Id. at 21.

[93] Id.

[94] Id.

[95] See Khaled El Emam, *Heuristics for De-Identifying Data*, SECURITY & PRIVACY, July/Aug. 2008, at 58; Khaled El Emam, *Risk-Based De-Identification of Health Data*, SECURITY & PRIVACY, May/June 2010, at 64 [hereinafter El Emam, *Risk-Based De-Identification*].

[96] El Emam, *Risk-Based De-Identification*, supra note 95, at 66.

countermeasures to those threats.[97]

B. PII 2.0 AND FAIR INFORMATION PRACTICES (FIPS)

In our reconceptualized notion of personal information, the key is to think about identification in terms of risk level. PII 2.0 conceives of identifiability as a continuum of risk rather than as a simple dichotomy. A clear way to demonstrate the functioning of this new approach is by considering the applicability of FIPs.

The basic toolkit of FIPs includes the following: (1) limits on information use; (2) limits on data collection, also termed data minimization; (3) limits on disclosure of personal information; (4) collection and use only of information that is accurate, relevant, and up-to-date (data quality principle); (5) notice, access, and correction rights for the individual; (6) the creation of processing systems that the concerned individual can know about and understand (transparent processing systems); and (7) security for personal data. When information refers to an *identified* person, all of the FIPs generally should apply.

As for the category of *identifiable*, it is not appropriate to treat such information as fully equivalent to identified. The information does not yet refer to a specific person and may never do so. Nonetheless, some protections are in order because there is a risk of linkage to a specific individual. The question then becomes, which of the FIPs should apply?

Full notice, access, and correction rights should *not* be granted to an affected individual simply because identifiable data about her are processed. For one thing, if the law created such interests, these obligations would decrease rather than increase privacy by requiring that all such data be associated with a specific person. This connection would be necessary in order to allow that individual to exercise her rights of notice, access, and correction. In this fashion, the law would create a vicious circle that could transform *identifiable* data into *identified* data. Indeed, Article 10 of the Proposed Data Protection Regulation provides that a data controller is not obligated to collect further information in order to identify the data subject for the mere purpose of complying with the proposed regulation.[98]

Moreover, limits on information use, data minimization, and restrictions on information disclosure should not be applied across the board to identifiable information. Such limits would be disproportionate to risks from data use and also would cripple socially productive uses of analytics that do not raise significant risks of individual privacy harms.[99] Some of these uses of analytics are consumer-oriented and some are not, but the benefit to the public is often clear.

As an example of a consumer-focused service, Google Flu Trends provides geographic information on the spread of the influenza virus based upon search queries entered into Google's search engine. Among nonconsumer uses of this analytics, analysis of large data sets

[97] See MICHAEL HOWARD & STEVE LIPNER, THE SECURITY DEVELOPMENT LIFECYCLE (2006) (discussing techniques for engineers to develop more secure software).

[98] Proposed Regulation, supra note 25, at art. 10.

[99] At the Article 29 Working Party of the European Union, there recently has been openness to a concept of proportionality in the use of information privacy law. See Article 29 Data Protection Working Party, *Opinion 3/2010 on the Principle of Accountability* at 3, WP 173 (July 13, 2010). The question remains as to how successful this concept can be in a system that treats identified and identifiable data as equivalents.

plays an important role in healthcare research, the management of physician performance and clinical metrics, data security, and fraud prevention. Regarding healthcare research, there has been a shift away from traditional clinical trials that follow specific patients toward informational research that analyzes large data and biological sample sets. The Institute of Medicine explains these new "information based" forms of inquiry as "the analysis of data and biological samples that were initially collected for diagnostic, treatment, or billing purposes, or that were collected as part of other research projects."[100] This technique, centered on analytics, is widely used today in categories of research including epidemiology, healthcare services, and public health services. These information-based forms of health research "have led to significant discoveries, the development of new therapies, and a remarkable improvement in health care and public health."[101]

While all FIPS should not apply to identifiable data, there are three that are applicable in this context. The key FIPs are those that concern data security, transparency, and data quality. Data security refers to the obligation to "protect against unauthorized access to and use, destruction, modification, or disclosure of personal information."[102] Identifiable information should be subject to data security principles. Recall that identifiable data are those for which a specific identification, while possible, is not a significantly probable event. Yet these data, unlike non-identifiable information, might be relatable to a person. Data security for identifiable information, as for identified information, should be commensurate with the nature of the information and the likely risks of disclosure.

As for transparency, this FIP calls for the creation of data processing systems that are open and understandable to affected individuals. Transparency also means that tracking or surveillance should not be done secretly. First, openness about information use allows for improved policies and law. As Louis Brandeis famously stated, "Sunlight is said to be the best of disinfectants; electric light the most efficient policeman."[103] Second, identifiable information can have great value. Transparency about the collection of identifiable information will serve to heighten awareness about data flows among all parties, both consumers and corporations. It will also improve the position of consumers who have preferences about the collection and further use of data—even should that information merely be identifiable.

Finally, data quality is a FIP that requires organizations to engage in good practices of information handling. This requirement depends on the purpose for which information is to be processed. In the context of *identified* data, for example, it means that the greater the potential harm to individuals, the more precise that the data and its processing must be. Some decisions matter more than others, and the stakes are low when the issue is whether or not one receives a coupon for a dollar discount on a case of seltzer. More accuracy is required for a data system that decides whether or not one receives a mortgage and calculates the interest rate associated with it. In contexts where the decision to be made about a person based on identified data is more important, or the harm to the person potentially greater, there must be higher requirements for data quality.

In the context of *identifiable* information, data quality also requires good practices of information handling. In particular, it requires that companies pay attention to the use and

[100] INST. OF MED. OF THE NAT'L ACADS., BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH 112 (Sharyl J. Nass et al. eds., 2009).
[101] Id. at 113.
[102] SOTTO, supra note 41, at 14-3.
[103] LOUIS D. BRANDEIS, OTHER PEOPLE'S MONEY AND HOW THE BANKERS USE IT 92 (1914).

processing of identifiable information by third parties. If information is non-identifiable, a company can publicly release it or permit third parties access to it without further obligations.

Identifiable information is capable of identification, even if this risk is not significantly probable. Thus, companies cannot merely release it or allow unmonitored access to it. Depending on the potential harm to individuals and the likely threat model, companies should also be required to use a "track and audit" model for some identifiable information. An example would be information used in health care research. Access to such data should be accompanied by obligations that travel with the information. Companies that handle identifiable information can structure these obligations by associating metadata, or information about information, with data sets.

C. PII 2.0 AND EU PRIVACY LAW

Our model of PII 2.0 has elements that are distinct from EU law and US law alike. As we have seen, the Article 29 Working Party would treat all information collected by a data controller as identifiable, and hence subject to full protection, so long as the ultimate intention is to link some of these data to individuals. Indeed, none of the information collected may ever be identified; this result is demonstrated by the hypothetical example of the Working Party concerning the "key-coded data." As a further matter, the Kuner example with the Verdi-loving physician shows how EU law lacks a requirement that the party who can link data to a specific individual be reasonably likely to obtain the information. In contrast, and as noted earlier, the Proposed Regulation's Recital 24 adopts a context-specific analysis and states that "location data … need not be considered as personal data in all circumstances."

The distinction with US law is also clear. US law tends only to protect identified information. State data breach notification law takes the approach; unless a last name, and first name, or first initial is present, the leakage of other information does not require the organization to inform the affected individual. The tautological approach is of even less help; it states that "personally identifiable information" is "information that identifies a person." This language seems to suggest that the identification must have already taken place and not depend on events reasonably likely to occur.

At the same time, PII 2.0 fuses US and EU privacy law by using concepts derived from each. From the US approach, PII 2.0 takes a more harm-based approach. Like US law, it finds that data about identified individuals should be given the most protection. From the EU approach, PII 2.0 recognizes that identifiable data still deserves protection and should be included within the definition of PII.

But is PII 2.0 compatible with the EU approach? Upon initial reflection, one might expect the answer to be "no." PII 2.0 provides only some of the FIPs to certain kinds of data that EU privacy law would protect with all of the FIPs. The EU approach also tends to be uniform regarding its FIPs, and PII 2.0 permits variations in protection. Thus, on the surface, PII 2.0 might appear to weaken EU privacy protection and contravene its goal of providing a uniform and high level of privacy protection to data in order to respect people's fundamental right to privacy.

In our view, however, PII 2.0 is fully compatible with the EU approach, consistent with its underlying philosophy, and furthers its goals effectively. The EU approach already diverges from uniformity when different levels of protection will better protect people's right to privacy.

As we have noted, the EU Data Protection Directive identifies a special category of data called "sensitive data" and provides it with stronger protections than other types of data.[104] The Proposed Regulation contains a similar category.[105] Our model of PII 2.0 leaves this category in place. As a larger point, the concept of sensitive data shows how EU already supports different categories of data with different levels of protection.

In addition, the Proposed Data Protection Regulation and the Article 29 Working Party have provided some indications that the full rigor of European Data Protection law need not apply to all types of personal data. At present, however, these do not represent the conventional wisdom, but demonstrate only some evidence of an evolution of the majority view. As an example, and as mentioned previously, the Proposed General Data Protection Regulation recognized that applying the Proposal's full requirements to identifiable data would create, at least at times, the perverse result of obligating organizations to collect more personal data in order to authenticate the data subjects. The drafters therefore wisely included Article 10, which provides that data controllers need not collect more personal information to identify the data subject for the mere purpose of complying with the proposed regulation.[106]

The logic of proposed Article 10 is impeccable – it recognizes that identifiable information *should not and cannot* be regulated the same as identifying information. Thus, while the Regulation does not specifically create two classes of personal data with differing requirements, Article 10 would permit such effects. Yet, Article 10 is no panacea. It is vague both as to the types of personal data to which it would apply and as to the provisions of the regulations for which a data controller need not comply if they have such information. Thus, while Article 10 is an important recognition that identified personal information should be regulated differently from identifiable information, it is an incomplete solution that PII 2.0 addresses more fully.

As for the Article 29 Party, its opinions have helped develop the EU's expansionist views of personal data. We have noted some of the shortcomings in this approach, such as redefining "identified" as decision-making based on specific characteristics or whether *some* parties in a database might be identified. Yet, in its 2011 opinion on geolocation data, the Article 29 Working Party found some information deserved a lighter set of FIPs because it posed a lesser privacy risk.[107]

To be sure, the Article 29 Working Party did broadly state that a data controller should treat "all data about WiFi routers as personal data."[108] Even when "in some cases the owner of the device currently cannot be identified without unreasonable effort," a WiFi access point should be viewed as personal data.[109] It reached this conclusion because the information can be indirectly identified in certain cases. This opinion of the Article 29 Working Party does not demonstrate flexibility in the definition of "personal data." Thus, its starting point was that a WiFi MAC address in combination with location information constituted "personal data." Yet, it also found that this information posed a "lesser threat to the privacy of the owners of these access points than the real-time tracking of the locations of smart mobile devices."[110] Due to

[104] Data Protection Directive, supra note 7, at art. 8.
[105] Proposed Regulation, supra note 25, at art. 9.
[106] Id. at art. 10.
[107] Article 29 Data Protection Working Party, *Opinion 13/2011 on Geolocation services on smart mobile devices* at 7 WP 185 (May 16, 2011).
[108] Id. at 11.
[109] Id.
[110] Id. at 17.

this "lesser threat," the Article 29 Working Party took some initial steps on the path to PII 2.0. It called for a less rigorous opt-out mechanism, rather than opt-in; a lighter notice requirement; and implied that access for the affected individual need not be provided if it would require collection of additional information would be collected to authenticate the WiFi access point owner**.**[111]

Thus, PII 2.0 is consistent with at least some existing strands in EU information privacy law. Most importantly, PII 2.0 enhances the protection of privacy. It creates an incentive for companies to keep information in the least identifiable form. If we abandon PII, or treat identified and identifiable information as equivalents, companies will be less willing to expend resources to keep data in the most de-identifiable state practicable, or, as the FTC proposes, to develop strong contracts with outside parties to keep shared information de-identified.

Beyond this incentive to keep data de-identified, PII 2.0 enhances privacy because administering certain FIPs requires that data be identified, and keeping data in identified format can create privacy risks. Providing individuals with access to their data, for example, requires that the information be kept in identified form. If data is not kept in this form, data processors would not know to whom to provide access. The problem is that by keeping the data in identified form, the privacy risks increase from a potential data security breach.

When a breach involves only *identifiable* data, in contrast, the harm that the information can cause to individuals is much less likely to occur. Harm can only occur when the party who obtains the data also knows how to identify it. Although identification of some data may be theoretically possible, individuals with unauthorized access to it may lack resources or the knowledge to do so. The idiosyncratic nature of certain supposed triumphs of re-identification and the technical complexity of identifying de-identified data in many circumstances is sometimes overlooked. For example, Daniel Barth-Jones has deflated the popular account of the easy re-identification from a medical data set in 1997 of William Weld, who was then Governor of Massachusetts.[112] Barth-Jones demonstrates both the re-identification rested on certain aspects of the population demographics of Cambridge, Massachusetts, where Weld was a registered voter, and the robust nature of the protections of the Privacy Rule, which the federal government issued pursuant to the Health Insurance Portability and Accountability Act.[113]

Moreover, keeping data in de-identified form prevents harms from inappropriate access by employees or others. It also makes it harder to engage in new uses of data, which is why the FTC seeks to have companies contractually prohibit downstream recipients from re-identification of shared data. Beyond these legal barriers, of course, the mere status of information in de-identified form creates barriers to identification through technological requirements, time, and cost.

Therefore, for the goal of protecting privacy, having data kept in identifiable rather than identified form is a significant plus. PII 2.0 encourages keeping data in this format. The EU approach to PII discourages keeping data merely identifiable. PII 2.0 would strengthen privacy protection in the EU and resolve some of the ambiguities currently in the EU Data Protection Directive.

[111] Id. at 16, 18.

[112] Daniel C. Barth-Jones, *The "Re-identification of Governor William Weld's Medical Information* (2012), at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

[113] Id.

IV. CONCLUSION

Despite what appears to be significant divergence between the concepts of personal data in the US and the EU, the differences between the two systems can be reconciled. PII 2.0 makes the US approach coherent, and it fits within the basic philosophy and principles of US privacy law. PII 2.0 also is consistent with the basic philosophy and principles of EU privacy law. Thus, PII 2.0 is the ideal starting point toward reconciling these divergent bodies of law.

PII 2.0 would enlarge the scope of many US privacy laws, but it would not impede data flows. It would not restrict the scope of EU privacy laws but would provide more tailored and nuanced protections. The differences between US and EU law are often viewed as insurmountable, but as we have argued in this Article, the differences can be effectively bridged while keeping the basic philosophy behind US and EU privacy law intact.