Harvard Journal of Law & Technology
Volume 37, Number 3 Symposium 2023

A SCIENTIFIC APPROACH TO TECH ACCOUNTABILITY

David Choffnes, Woodrow Hartzog,** Scott Jordan,****
*Athina Markopoulou**** & Zubair Shafiq******

ABSTRACT

The Federal Trade Commission can only do so much to hold tech companies accountable. Enforcement agencies and the people they protect need help. One problem is that the inner workings of large organizations and complex algorithmically driven systems remain obscure and opaque while their privacy representations are voluminous and vague. In this Essay, we propose a scientific approach to tech accountability, where academic researchers can play a larger role in privacy policy. This approach involves surfacing a company's privacy representations and statements, as well as measuring the actual behavior of their systems with respect to algorithms, user interfaces, and data processing.

We build upon our experience as a multi-disciplinary group of researchers trained in computer science, engineering, and law to explore how researchers can support the movement for tech accountability. In addition to detailing how researchers can surface a company's privacy representations and measure the behavior of tech systems, we explore how to use scientific results for greater accountability, such as going public, working with regulators, filing Unfair, Deceptive, or Abusive Acts or Practice ("UDAAP") complaints and lawsuits, and taking advantage of data subject rights. We draw from our own research to demonstrate how this approach can be helpful, such as in uncovering significant discrepancies between privacy representations of tech companies and the actual behavior of their systems and devices. We conclude by calling for a more robust and long-term collaboration between researchers and regulators.

---

* Khoury College of Computer Sciences, Northeastern University.

** Professor of Law, Boston University; Faculty Associate, Berkman Klein Center for Internet & Society at Harvard University; Affiliate Scholar, Stanford Law School Center for Internet & Society.

*** Department of Computer Science, University of California, Irvine.

**** Department of Electrical Engineering and Computer Science, University of California, Irvine.

***** Department of Computer Science, University of California, Davis. Other acknowledgements: This work was supported by ProperData, a Secure and Trustworthy Cyberspace ("SaTC") Frontiers Project funded by National Science Foundation Awards 1956393, 1955227, 1956435, and 2103439. The authors would like to thank Janelle Robins and Kabbas Azhar for their excellent research assistance.

TABLE OF CONTENTS

I. INTRODUCTION

U.S. regulators like the Federal Trade Commission ("FTC") have their hands full keeping organizations accountable for their privacy and data practices. They need help. Regulators have limited powers and resources. Most can only bring a handful of enforcement actions at a time. And people suffering privacy violations have only a few meaningful options to privately bring claims against tech companies. Meanwhile, the inner workings of large organizations and complex, algorithmically-driven systems remain obscure and opaque while the privacy representations of those organizations are voluminous and vague. Regulators that police unfair and deceptive trade practices, like the FTC, need assistance in identifying who is claiming what, how systems actually work, and whether there is any discrepancy between them.

In this Essay, we argue that academic researchers are well suited to help identify and understand privacy violations, and therefore their research contributions should be more explicitly considered by lawmakers. We call this a "scientific approach to tech accountability" and propose this approach as a way to aid both lawmakers with limited resources and enforcement authority, and people with limited ability to protect their privacy. Regulators are struggling because the three major rules of privacy law in the United States, "Follow the Fair Information

Practices," "Do Not Harm," and "Do Not Lie," all depend upon an elusive transparency and access to information systems. We draw upon our own research to demonstrate how a scientific approach to tech accountability might help regulators. As a multi-disciplinary group of researchers from the fields of computer science, engineering, and law, we have experience creating and interrogating data that could prove useful to regulators. If lawmakers were to structurally support a scientific approach to tech accountability, we envision academic researchers playing a larger role in privacy policy, working with public interest technologists and regulators to better enforce the privacy rules that we have and highlight the need for reform. Although we focus on U.S. law in this Essay, our approach applies in any context with similar data privacy frameworks and institutional commitments.

Our argument proceeds in three parts. In Part II, we highlight privacy's accountability gap. For better or worse (probably worse), data privacy rules' enforceability in the United States is mostly dependent upon the representations made by tech companies and a clear understanding of how their complex, opaque systems affect consumers. Unfortunately, access to these systems is difficult to come by and the representations made by companies are hidden, incomplete, vague, and voluminous. In this Part, we explore how the three main rules of privacy — "follow the Fair Information Practices," "do not harm," and "do not lie" — place great weight on representations and access. We also explore the structural, political, and practical limitations placed on regulators like the FTC. Putting aside questions about whether our current privacy rules are sufficient (they are not), we argue that the law's substantive and structural shortcomings prevent regulators from meaningfully enforcing existing rules. For example, companies routinely break privacy promises, harm consumers, and fail to follow the Fair Information Practices ("FIPs") without meaningful regulatory pushback.

In Part III, we outline our scientific approach to tech accountability. First, we propose that researchers can support the enforcement of privacy laws by some combination of (1) surfacing a company's privacy representations; and (2) measuring the actual behavior of a company's systems with respect to their algorithms, user interfaces, and processing of data. We explore how to find and understand a company's privacy representations through automated analyses of statements made by the company, as well as through leveraging data subject access rights. We describe multiple methods we have used to measure the actual behavior of user systems, including network traffic and ad targeting analysis. In our approach, we apply the scientific method of exploratory inquiry, hypothesis development, rigorous testing, and sound data analysis to prove or contradict our hypotheses in a way that stands up to scrutiny. As part of our approach, we consider not only the individual

company's representations, implementations, and practices, but we also generalize to different companies, multiple modalities, and future technologies and develop systematic methods that are broadly applicable.

In Part IV, we explore how to use the scientific approach for greater accountability. First, we discuss how to use existing policy levers, including going public, working with regulators filing UDAAP complaints, filing lawsuits, and taking advantage of data subject rights to keep tech accountable. We draw from our own empirical research to demonstrate how this approach might be helpful. For example, the application of our scientific approach to Amazon's smart speaker ecosystem uncovered discrepancies between the privacy representations and actual behavior of Amazon and third-party skills.[1] Among other things, our work has served as the basis of a consumer class action lawsuit.[2] Second, we explore what new rules would best support collaborations between regulators and researchers for applying the scientific approach to tech accountability. We conclude this Essay by applauding the FTC's newly launched Office of Technology and exploring a more robust and long-term collaboration between researchers and regulators.

II. PRIVACY'S ACCOUNTABILITY GAP

Privacy law is expansive, but its commercial core can be distilled down to three major rules, all of which depend upon companies being transparent and clear about their privacy practices and how their systems work.

The first rule of privacy is "Follow the Fair Information Practices," or FIPs.[3] The FIPs serve as the world's preeminent privacy and data protection model. Originating from a 1970s report from the U.S. Department of Health, Education, and Welfare, these influential principles are considered the gold standard for privacy.[4] The Organisation for Economic Co-operation and Development later revised the principles in a widely recognized document, which now serves as the foundation for privacy regulatory schemes and public policy.[5] The FIPs shape U.S. privacy statutes, such as the Health Insurance Portability and

1. Skills are Alexa's (Amazon's smart speaker's) equivalent of apps. Umar Iqbal, Pouneh Nikkah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel Dubois et al., *Tracking, Profiling, and Ad Targeting in the Alexa Echo Smart Speaker Ecosystem*, 2023 PROC. ACM INTERNET MEASUREMENT CONF. 569, 569 [hereinafter *Echos*].

2. *See* Gray v. Amazon.com, Inc., No. 2:22-cv-800, 2023 WL 1068513 (W.D. Wash. Jan. 27, 2023).

3. *See* WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 59 (2018).

4. *See* U.S. DEP'T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS: REPORT OF THE SECRETARY'S ADVISORY COMMITTEE ON AUTOMATED DATA SYSTEMS at xxvii–xxviii (1973).

5. *See* ORG. FOR ECON. COOP. & DEV., GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA 3–4 (1980).

Accountability Act of 1996[6] ("HIPAA"), and are utilized by the FTC to police unfair and deceptive trade practices.[7] The FIPs boil down to transparency and safety principles such as notice and choice, access and correction, and security.[8] The FIPs depend upon comprehensible company representations and systems.[9]

The second privacy rule is "Do Not Harm," which requires that companies not injure people through their data practices and system design.[10] This rule is reflected in the laws and regulations that prevent unfair trade practices and require companies to take care in protecting people's data. For instance, in order to bring a claim against a data collector, people usually need to show that they have suffered some kind of financial harm.[11] The Federal Trade Commission Act[12] defines an unfair practice as one that causes or is likely to cause significant harm to consumers that they cannot avoid and that is not outweighed by benefits.[13] The EU General Data Protection Regulation[14] ("GDPR") allows data subjects to seek remedies if they have suffered damage as a result of illegal data processing.[15] The privacy torts are all anchored by a "do not harm" mentality.[16] However, for the "do not harm" rule to be effective, people need to understand how technology works and how it is intended to be used. Otherwise, they may be more likely to use it in ways that leave them vulnerable to harm. Regulators also need to show how a company's wrongful behavior has caused harm to people, which can be difficult if they do not have a good understanding of how the company's systems work.

Finally, and most importantly for our approach, the third commitment of information privacy law is "Do Not Lie."[17] Privacy law can abide all kinds of dubious behavior, but one of its foundational rules prohibits lies and misrepresentations. Rules like the FTC's prohibition on deceptive trade practices mandate that companies be truthful in their

---

6. Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of the U.S. Code).

7. *See* Woodrow Hartzog, *The Inadequate, Invaluable Fair Information Practices*, 76 MD. L. REV. 952, 960 (2017).

8. *See* HARTZOG, *supra* note 3, at 60, 64.

9. *See generally* Hartzog, *supra* note 7.

10. *See* HARTZOG, *supra* note 3, at 70–72.

11. Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 TEX. L. REV. 737, 741–42 (2018).

12. 15 U.S.C. §§ 41–58.

13. *Id.* § 45(n).

14. Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) [hereinafter GDPR].

15. *Id.* art. 77.

16. *See* Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. REV. 793, 809 (2022).

17. *See* HARTZOG, *supra* note 3, at 67.

privacy policies, marketing, and contracts.[18] Privacy laws, including the Gramm-Leach-Bliley Act, HIPAA, Health Breach Notification Rule, and state data-breach notification laws, require disclosure of privacy-related items in the form of privacy policies.[19] In the 1990s, as people began to use the Internet for commercial activities and personal data could be more easily gathered, the technology industry favored self-regulation with the "notice and choice" approach.[20] This largely involved companies including privacy policies on their websites and giving users the option to opt-out by simply not using the service or browsing the website. Privacy policies are now a standard part of websites and apps, appearing as dense and often unreadable terms of use agreements. These agreements are typically considered binding contracts, but they differ from classic contracts in that they are usually non-negotiable.[21] And of course, contracts, privacy policies, and marketing can all induce reliance, so it is important that they are truthful.[22]

To determine if a company's privacy claims are true, they must be comprehensible. That leads us to privacy's accountability gap: Our rules for tech accountability depend upon transparency and clarity, and yet accountability is stymied because information systems are opaque and the representations made by companies are incomprehensible. More specifically, many products that collect and use data gathered from consumers do so via closed systems (e.g., Google Analytics, Amazon Echo devices) that are hidden from public view (i.e., independent parties cannot access device hardware, view software source code, access the code and data analyzed at servers, or inspect network traffic contents) where openness is often equated with disclosure of trade secrets.[23] In addition, representations made by companies entail the fine print of lengthy legal documents, often with further linked documents, contradictory statements, and vague disclosures that frustrate attempts

18. Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 628 (2014).

19. Gramm-Leach-Bliley Act of 1999, 15 U.S.C. § 6803; Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936; 45 C.F.R. § 164.530(i) (requiring HIPAA covered entities to designate privacy official to develop and implement "policies and procedures" of the entity); Health Breach Notification Rule, 16 C.F.R. § 318.3 (2023); CAL. CIV. CODE § 1798.29 (West 2023); *see, e.g.*, MASS. GEN. LAWS ch. 93H, § 3 (2023).

20. Solove et al., *supra* note 18, at 592.

21. Woodrow Hartzog, *Website Design as Contract*, 60 AM. U. L. REV. 1635, 1640 (2011); Woodrow Hartzog, *Promises and Privacy: Promissory Estoppel and Confidential Disclosure in Online Communities*, 82 TEMP. L. REV. 891, 921 (2009); Woodrow Hartzog, *The New Price to Play: Are Passive Online Media Users Bound by Terms of Use?*, 15 COMMC'N L. & POL'Y 405, 413 (2010).

22. *See* Hartzog, *Website Design*, *supra* note 21, at 1661.

23. *See, e.g.*, Rebecca Wexler, *It's Time to End the Trade Secret Evidentiary Privilege Among Forensic Algorithm Vendors*, BROOKINGS (July 13, 2021), https://www.brookings.edu/articles/its-time-to-end-the-trade-secret-evidentiary-privilege-among-forensic-algorithm-vendors [https://perma.cc/3UDB-BVGE].

to understand an individual's privacy risks. Privacy's accountability gap matters for consumers, for government officials, for auditors, and even for those within the companies themselves. In this Part, we explore how incomprehensible representations and opaque systems plague accountability efforts.

A. Privacy Representations Are Incomprehensible

In theory, regulators have information available that can help keep companies honest. Virtually every website has a privacy policy detailing their privacy practices, most of which follow a similar format and tone, using similar subsections and sometimes even the same language. While many of these disclosures are voluntary, privacy representations are often required by law.[24] In California and Europe, statutes go beyond requiring only accurate privacy disclosures to requiring specific types of disclosures: both Europe's GDPR and California's California Consumer Privacy Act of 2018[25] ("CCPA") require that a business's privacy policy disclose the categories of personal information collected, used, and shared, and the purposes for collecting, using, and sharing personal information.[26]

However, these disclosures are often incomprehensible because they are inconsistent, vague, and far too numerous. They do not enable people to make informed decisions about their use of services and applications and are often insufficient for regulatory bodies to determine the accuracy of disclosures.[27]

First, privacy representations are wildly different from each other, making comparisons hard and identifying baseline levels of specificity for disclosures even harder. In the absence of a statutory or regulatory requirement that privacy policies adhere to standardized definitions of "personal information" or "sensitive personal information," privacy policies often define personal information or sensitive personal information differently than the GDPR or the CCPA do, or they fail to define personal information whatsoever.[28] Consequently, privacy policies' definitions of personal information usually end up being far too narrow,

---

24. *See* sources cited *supra* note 19.

25. CAL. CIV. CODE §§ 1798.100–1798.199 (West 2023).

26. GDPR, *supra* note 14, arts. 13, 14; CAL. CIV. CODE § 1798.130(a)(5) (West 2023). *See generally* Scott Jordan, *Strengths and Weaknesses of Notice and Consent Requirements Under the GDPR, the CCPA/CPRA, and the FCC Broadband Privacy Order*, 40 CARDOZO ARTS & ENT. L.J. 113, 134–136,138–140, 143–146 [hereinafter *Strengths*].

27. Solove et al., *supra* note 18, at 634, 667; *see* HARTZOG, *supra* note 3, at 64, 141.

28. Compare the various terms used to describe personal information in privacy policies, Scott Jordan, Siddharth Narasimhan & Jina Hong, *Deficiencies in the Disclosures of Privacy Policy and in User Choice*, 34 LOY. CONSUMER L. REV. 408, 429 n.123, 435 n.157, 445 n.213, 454 n.270, 466 n.340 [hereinafter *Deficiencies*], with the definition of "personal information" in the CCPA, CAL. CIV. CODE § 1798.140(v)(1) (West 2023), and the definition of "personal data" in the GDPR, GDPR, *supra* note 14, art. 4(1).

excluding information that does not itself identify a person but which can be used to reasonably identify a person and information paired with a device identifier which can be reasonably linked to a person.[29]

Similarly, in the absence of a statutory or regulatory requirement that privacy policies adhere to a standardized definition of "de-identified information," privacy policies often define de-identified information differently than the CCPA or fail to define de-identified information whatsoever.[30] Consequently, privacy policies' descriptions of anonymous and de-identified information are far too broad, including information paired with advertising identifiers that the computer science literature has repeatedly demonstrated is reasonably linkable.[31]

Second, companies' privacy representations are vague. A common format for disclosures in privacy policies is to separate disclosures about what information a company collects, how personal information is used, and how the company shares personal information.[32] As a result of these fragmented disclosures, privacy representations typically fail to indicate how specific kinds of information are used or shared.[33] Because privacy policies' disclosures of the uses of personal information are usually disconnected from their disclosures about the types of personal information collected, we are usually unable to determine which types of information are used for which purposes.[34]

For example, we generally cannot determine whether location or web browsing history is used solely for functional purposes or also for advertising.[35] It is unclear whether the CCPA and the GDPR require a privacy policy to disclose the purpose of collecting information for each category of personal information collected.[36] Lawmakers can be blamed for some of this confusion. For example, the CCPA requires a privacy policy to disclose the purpose of sharing for each category of personal information shared.[37] However, because privacy policies' disclosures of *sharing* of personal information are usually presented in a different section of the policy than their disclosures about the types of

29. *Deficiencies*, *supra* note 28, at 429 n.123, 435 n.157, 445 n.213, 454 n.270, 466 n.340. Examples include web browsing histories and information paired with Apple or Android advertising identifiers.

30. Compare the various terms used to describe personal information in privacy policies, *id.* at 429 n.123, 435 n.157, 445 n.213, 454 n.270, 466 n.340, with the definition of "deidentified information" in the CCPA, CAL. CIV. CODE § 1798.140(m) (West 2023).

31. *Deficiencies*, *supra* note 28, at 429 n.123, 435 n.157, 445 n.213, 454 n.270, 466 n.340.

32. *Id.* at 427–75.

33. *Id.*

34. *Id.* at 429–31, 435–36, 444–46, 454–57, 466–67, 472–73.

35. *Id.* at 475–81.

36. *Strengths*, *supra* note 26, at 139–40.

37. *Id.* at 143–46. It is unclear whether the GDPR has similar requirements.

personal information *collected*, we are usually unable to determine which types of information are shared.[38]

Finally, there are far too many privacy policies for the FTC to review for accuracy or for the new California Privacy Protection Agency to review for compliance with the CCPA. Regulatory agencies often rely on a combination of internal reviews of privacy policies, formal complaints submitted to the agency, and investigation by stakeholders and media to raise red flags about possible violations.[39] However, even the combination of these triggers can review only a small fraction of the privacy policies on the Internet.[40] Below, we call for the use of automated processes to examine privacy policies and to raise red flags that regulatory agencies can then examine to determine whether violations have occurred.

B. Information Systems Are Opaque

Independent of the nature of representations made by companies, today's information systems are generally so opaque that there is no reasonable way to independently verify their claims and keep companies in check. Specifically, technology is often a "black box" where the hardware, software, and data transmission entailed in online systems are kept secret and hidden from independent parties.[41] As a result, regulators today have no choice but to take a position of trusting companies by default and can only take action retroactively when flagrant harms are publicized.[42] Unfortunately, online systems have given us no reason to trust these companies by default, and ad hoc approaches to enforcement leave far too many harms on the table.

It does not have to be this way. We argue that a scientific approach can address many of the challenges in this area. Specifically, such an approach can enable systematic, repeatable, automated, and rigorous evaluations of online systems. In turn, this can change the conversation from "just trust us" to "trust, but verify" by default and enable the

38. *Deficiencies*, *supra* note 28, at 431–32, 436–37, 446–47, 457–61, 467–69, 474.

39. Solove et al., *supra* note 18, at 609.

40. There are an estimated 1.13 billion websites on the Internet. Kathy Haan, *Top Website Statistics for 2023*, FORBES ADVISOR (Feb. 14, 2023), https://www.forbes.com/advisor/business/software/website-statistics [https://perma.cc/66YD-CTP4].

41. Kashmir Hill, *These Academics Spent the Last Year Testing Whether Your Phone Is Secretly Listening to You*, GIZMODO (July 3, 2018), https://gizmodo.com/these-academics-spent-the-last-year-testing-whether-you-1826961188 [https://perma.cc/X2CX-2ZZQ]; Jess Weatherbed, *This Site Exposes the Creepy Things In-App Browsers from TikTok and Instagram Might Track*, VERGE (Aug. 19, 2022), https://www.theverge.com/2022/8/19/23312725/in-app-browser-tracking-facebook-instagram-privacy-tool [https://perma.cc/W5Q7-VPZ7]; *see, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 3 (2015).

42. *See, e.g.*, Nicholas Confessore, *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*, N.Y. TIMES (Apr. 4, 2018), https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html [https://perma.cc/2ZPN-XTDD].

identification and remediation of harms prospectively instead of reactively — even as information systems change over time.

Transparency and accountability are of interest to all parties involved in this space. First, users want to understand how their data is treated and what their rights and options are. Second, regulators want to hold companies accountable and have a systematic and ideally automated way to audit data collection and use practices, as opposed to relying on anecdotal evidence and ad hoc findings. Policy makers also want to understand the trends and current practices regarding data so they can update the privacy laws and regulations. Third, even when companies want to comply with privacy laws, it is challenging to do so.[43] Developers often do not fully understand the information flow in their own systems, due to their complexity, "time to market" pressure, and use of third-party software or hardware with their own opaque data practices.[44]

For example, privacy statutes do not adequately address when a first party uses a software library provided by a third party. Third party software libraries commonly allow the third party itself to collect, use, and share personal information from the consumer.[45] However, privacy statutes rarely recognize this situation. They often define the first party as the party with whom a consumer intentionally interacts and a third party as a party with whom a consumer does not intentionally interact.[46] Both the CCPA and the GDPR hold the first party responsible for the activities of a third party when it outsources tasks to that third party under a contract.[47] However, they are less clear about the responsibilities of the first party when it allows a third party to collect information from the consumer outside such contracts.[48]

Ad tech also relies on complex and opaque systems and technologies (e.g., real-time bidding) to enable advertisers to programmatically

---

43. *See* Sam Biddle, *Facebook Engineers: We Have No Idea Where We Keep All Your Personal Data*, INTERCEPT (Sept. 7, 2022), https://theintercept.com/2022/09/07/facebook-personal-data-no-accountability/ [https://perma.cc/A4R4-ZNKD].

44. Alina Stöver, Nina Gerber, Henning Pridöhl, Max Maass, Sebastian Bretthauer, Indra Spiecker gen. Döhmann et al., *How Website Owners Face Privacy Issues: Thematic Analysis of Responses from a Covert Notification Study Reveals Diverse Circumstances and Challenges*, 2023 PROC. ON PRIV. ENHANCING TECHS. 251, 251; Nikita Samarin, Shayna Kothari, Zaina Siyed, Oscar Bjorkman, Reena Yuan, Primal Wijesekera et al., *Lessons in VCR Repair: Compliance of Android App Developers with the California Consumer Privacy Act (CCPA)*, 2023 PROC. ON PRIV. ENHANCING TECHS. 103, 115 (2023); Dominik Breitenbach, Ivan Homoliak, Yan Lin Aung, Nils Ole Tippenhauer & Yuval Elovici, *Hades-IoT: A Practical Host-Based Anomaly Detection System for IoT-Devices (Extended Version)*, 9 IEEE INTERNET THINGS J. 9640, 9640–41 (2022).

45. Scott Jordan, *A Proposal for Notice and Choice Requirements of a New Consumer Privacy Law*, 74 FED. COMMC'NS L.J. 251, 318 (2022) [hereinafter *Proposal*].

46. *Id.*

47. *Id.* at 315–18.

48. *See Strengths*, *supra* note 26, at 140–42.

target ads to consumers based on their browsing activity.[49] Ad tech involves complex interactions between multiple parties such as publishers, advertisers, ad exchanges, ad networks, and data brokers.[50] There are tens of thousands of companies involved in ad tech, and it is not uncommon for dozens of entities to be involved in a single ad tech transaction.[51] These data sourcing and sharing relationships between different companies are not transparent to consumers. Moreover, the use of automated decision-making algorithms, such as machine learning, makes it difficult to understand how ads are targeted and why.[52] This complexity and lack of transparency makes it difficult to understand how ad tech works — specifically, the collection, sharing, and processing of personal data by different companies in service of targeted advertising.

As an example of a complex system that collects data, consider the Internet of Things ("IoT"). An increasing number of smart interconnected objects are becoming affordable, popular, and rich in functionality, with up to twenty-nine billion devices expected to be deployed globally by 2027.[53] While these devices enable a wide range of societal benefits including health, safety, accessibility, and sustainability, they also present important privacy challenges. For example, smart TVs have been caught inferring and selling consumer viewing habits without consent,[54] and smart speakers profile consumers and use this data for advertising.[55] The troves of user data to which IoT devices have access from their sensors, their typical always-on nature, their unrestricted network access, the delegation of some of their computation to the cloud, and the fact that they are often closed platforms — meaning that they provide no easy audit access on how they work internally and on what information they propagate to other parties on encrypted connections (which are the vast majority)[56] — all create new privacy

---

49. *See* Andrew McStay, *Micro-Moments, Liquidity, Intimacy and Automation: Developments in Programmatic Ad-tech*, *in* COMMERCIAL COMMUNICATION IN THE DIGITAL AGE: INFORMATION OR DISINFORMATION? 143, 143 (Gabriele Siegert, M. Bjørn von Rimscha & Stephanie Grubenmann eds., 2017).

50. John Cook, Rishab Nithyanand & Zubair Shafiq, *Inferring Tracker-Advertiser Relationships in the Online Advertising Ecosystem Using Header Bidding*, 2020 PROC. ON PRIV. ENHANCING TECHS. 65.

51. *See id.* at 144.

52. *See* Davide Castelvecchi, *Can We Open the Black Box of AI?*, NATURE (Oct. 5, 2016), https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 [https://perma.cc/U68E-SSN5].

53. Satyajit Sinha, *State of IoT 2023*, IOT ANALYTICS (May 24, 2023), https://iot-analytics.com/number-connected-iot-devices [https://perma.cc/9B4M-QBL8].

54. Lesley Fair, *What Vizio Was Doing Behind the TV Screen*, FTC (Feb. 6, 2017), https://www.ftc.gov/business-guidance/blog/2017/02/what-vizio-was-doing-behind-tv-screen [https://perma.cc/8NH9-SVDF].

55. *Echos*, *supra* note 1.

56. Jingjing Ren, Daniel J. Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun & Hamed Haddadi, *Information Exposure from Consumer IoT Devices: A*

concerns. In an age where data is increasingly considered a commodity, IoT offers a very large surface for abuse with little possibility of knowing what data is collected, used, and sent to whom.

III. SCIENTIFIC APPROACH TO TECH ACCOUNTABILITY

We believe that a more scientific approach will be useful in helping keep tech companies accountable for their representations and services. Such a scientific approach should apply the scientific method of exploratory inquiry, hypothesis development, rigorous testing using systematic methodologies, and sound data analysis to prove or contradict our hypotheses. Such methodologies should ideally be automated, applicable at scale and across ecosystems, and repeatable over time. This is in contrast to one-off, anecdotal findings that rely on manual inspection. Although there is no one-size-fits-all auditing methodology for all questions of interest (e.g., Does company X collect my personal data? Which data? How do they use it?) or under all constraints (e.g., Does the auditor have access to the internals of the system under audit or is it a black box?), it is worth developing scientific approaches whenever possible.

In this Part, we explore how researchers can support the enforcement of privacy laws by some combination of surfacing a company's privacy representations and statements and measuring the actual behavior of a company's systems with respect to their algorithms, user interfaces, and data processing.

A. Surfacing a Company's Privacy Representations

First, we describe how researchers can help surface a company's privacy representations. These representations include statements made to the public, directly to data subjects, and in privacy policies. Regulators require companies to, at the very least, be honest, so statements on the record provide opportunities for accountability.[57]

1. Automating Privacy Policy Analysis

One important and legally binding representation of a company's practices is its privacy policy. Privacy policies have received considerable attention from the research community, which has made significant progress in automating the analysis of companies' privacy policies

Multidimensional, Network-Informed Measurement Approach, 2019 PROC. INTERNET MEASUREMENT CONF. 267, 267.

57. *See supra* notes 17–19 and accompanying text.

using Natural Language Processing ("NLP") in recent and ongoing work.[58]

Generally speaking, NLP-based policy analysis takes as input the text of a privacy policy, identifies statements related to how data is handled, extracts information (e.g., which entity; whether the data is collected, used, and/or shared; what data types; for what purposes; with or without consent; etc.), and represents it in pre-defined data structures (typically tuples or knowledge graphs)[59] connecting the aforementioned information. An NLP policy analyzer performs well if it has high coverage (i.e., identifies many collection statements) and is accurate (i.e., has few false positives and few false negatives). NLP privacy policy analyzers enable several applications. They can identify common patterns in texts across several different privacy policies and enable summarization.[60] Researchers can also use the precise representation of collection statements to detect contradictions within a privacy policy itself or between a privacy policy and external sources (e.g., privacy laws or other privacy policies).[61] Finally, researchers can check whether a privacy policy is consistent with the actual handling of data by the corresponding system.[62] Next, we describe some representative examples of NLP-based privacy policy analysis.

PolicyLint[63] was the first to provide an NLP pipeline that takes as input a sentence and outputs a collection statement.[64] For example, from the sentence "We may collect your email address and share it for advertising purposes," PolicyLint extracts as the collection statement (entity = "we", action = "collect", data type = "email address").[65] PolicyLint extracts collection statements from different sentences, considered in isolation from each other and represented as a list of independent tuples; each tuple is a list of ordered values (entity, action,

58. *See, e.g.*, Athina Markopoulou, Rahmadi Trimananda & Hao Cui, A CI-based Auditing Framework for Data Collection Practices, ARXIV (Mar. 30, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2303.17740 [https://perma.cc/CR9P-W97H].

59. Hao Cui, Rahmadi Trimananda, Athina Markopoulou & Scott Jordan, *PoliGraph: Automated Privacy Policy Analysis Using Knowledge Graphs*, 32 PROC. USENIX SEC. SYMP. 1037, 1038 (2023); Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi & Junbum Shin, *Consistency Analysis of Data-Usage Purposes in Mobile Apps*, 2021 PROC. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 2824, 2824; Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves et al., *PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play*, 28 PROC. USENIX SEC. SYMP. 585, 586 (2019).

60. Cui et al., *supra* note 59, at 1039.

61. *Id.*

62. Bui et al., *supra* note 59, at 2824.

63. Andow et al., *supra* note 59, at 585.

64. More generally, PolicyLint takes the app's entire privacy policy text, parses sentences, performs NLP techniques, and eventually extracts data collection statements defined as tuples of the form $P = \langle$app, data type, entity$\rangle$; app is the sender and entity is the recipient organization/entity performing an action ("collect" or "not collect") on the data type, and outputs: $P = \langle sender = platform/app, recipient = entity, data\ type \rangle$. *Id.*

65. *Id.* at 589.

data type).[66] In our own recent work, Poligraph,[67] we developed a privacy policy analyzer based on knowledge graphs, instead of tuples, to analyze the entire text of a privacy policy and capture relations between different sentences. Both tools entail a notion of ontology that captures subsumption relations between general and specific terms in a privacy policy — for example, that "email address" is a special case of "personal identifier."[68]

PoliCheck[69] builds on the collection statement tuples (entity, action, data type) extracted by PolicyLint from the privacy policy text and compares them to data flows observed in the network traffic generated by the corresponding software. It analyzes the consistency of the two and classifies the disclosures made in a privacy policy as clear (if the data flow exactly matches a collection statement), vague (if the data flow matches a collection statement in broader terms), omitted (if there is no collection statement corresponding to the data flow), ambiguous (if there are contradicting collection statements about a data flow), or incorrect (if there is a data flow for which the collection statement states otherwise).[70]

The purpose for collecting, using, and/or sharing personal information can also be automatically extracted from privacy policies. Polisis was one of the first NLP tools to extract purposes by classifying entire text segments.[71] MobiPurpose infers data collection purposes of mobile apps using network traffic and app features (e.g., URL paths, app metadata, domain name, etc.).[72] PurPliance automates the inference of data collection purposes introduced in MobiPurpose, extracts purposes from the privacy policy, and checks the consistency of policy text and system behavior, taking into account the consistency of purposes as well.[73]

---

66. *Id.*

67. Cui et al., *supra* note 59, at 1037.

68. *Id.* at 1040. PoliGraph makes a clear distinction between local and global ontologies to capture the context of individual privacy policies, application domains, and privacy laws. *Id.*

69. Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh et al., *Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck*, 29 PROC. USENIX SEC. SYMP. 985, 989 (2020).

70. *Id*. at 987–88.

71. Hamza Harkous, Kassem Fawaz, Remi Lebret, Florian Schaub, Kang G. Shin & Karl Aberer, *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning*, 27 PROC. USENIX SEC. SYMP. 531, 531–32 (2018).

72. Haojian Jin, Minyi Liu, Kevin Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson et al., *Why Are They Collecting My Data?: Inferring the Purposes of Network Traffic in Mobile Apps*, PROC. ACM ON INTERACTIVE, MOBILE, WEARABLE & UBIQUITOUS TECHS., 2018, at 1, 4.

73. Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi & Junbum Shin, *Consistency Analysis of Data-Usage Purposes in Mobile Apps*, 2021 PROC. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 2824, 2825.

In terms of application domains, the aforementioned NLP tools were originally developed for mobile apps[74] and later applied to Alexa skills (first here[75] and recently by us[76]), as well as to Oculus VR and apps.[77] With the advent of ChatGPT, this automated privacy policy analysis will likely be further accelerated, may become accessible to non-experts, and may also lead to custom language models such as PolicyGPT[78] specifically for privacy policies.

Overall, we believe that this type of analysis will be an important tool for auditing companies' representations in an automated way and comparing them to the corresponding system behavior.

2. Data Subject Access Rights ("DSARs")

Another legally binding representation of a company's practices concerns its response to data subject access rights ("DSARs"). The GDPR and the CCPA grant individuals certain rights, including the rights to know and delete their personal information collected by companies.[79] More specifically, DSARs also enable individuals to identify whether their data is being used for purposes they did not consent to, and if their data is being shared with third parties.[80] Violations of DSARs can serve as evidence in legal proceedings to hold companies accountable.[81] If a company provides incomplete or inaccurate information that does not match its actual behavior, consumers may have legal recourse to seek damages or other remedies.[82]

---

74. *See* Xiaoyin Wang, Xue Qin, Mitra Bokaei Hosseini, Rocky Slavin, Travis D. Breaux & Jianwei Niu, *GUILeak: Tracing Privacy Policy Claims on User Input Data for Android Applications*, 40 PROC. INT'L CONF. ON SOFTWARE ENG'G 37, 39 (2018); Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub et al., *Automated Analysis of Privacy Requirements for Mobile Apps*, NETWORK & DISTRIB. SYS. SEC. SYMP., Feb. 2017, at 1, 1.

75. Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das & William Enck, *Hey Alexa, Is This Skill Safe?: Taking a Closer Look at the Alexa Skill Ecosystem*, NETWORK & DISTRIB. SYS. SEC. SYMP., Feb. 2021, at 1, 1.

76. *Echos*, *supra* note 1.

77. Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anatasia Shuba & Athina Markopoulou, *OVRseen: Auditing Network Traffic and Privacy Policies in Oculus VR*, 31 PROC. USENIX SEC. SYMP. 3789, 3789 (2022).

78. Chenhao Tang, Zhengliang Liu, Chong Ma, Zihao Wu, Yiwei Li, Wei Liu et al., PolicyGPT: Automated Analysis of Privacy Policies with Large Language Models (Sept. 20, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2309.10238.pdf [https://perma.cc/H7XJ-D7VS].

79. GDPR, *supra* note 14, arts. 15–17; CAL. CIV. CODE §§ 1798.105, 1798.106, 1798.110 (West 2023).

80. *See, e.g.*, CAL. CIV. CODE §§ 1798.110, 1798.115 (West 2023) (giving California consumers a right to know what information is being collected as well as a right to know to whom it is being sold).

81. *See* Margot E. Kaminski, *The Case for Data Privacy Rights (Or 'Please, a Little Optimism')*, 97 NOTRE DAME L. REV. REFLECTION 385, 398 (2022).

82. *See id.* at 397.

It is worth noting that CCPA includes a provision for "authorized agents," which enables a consumer to authorize a third-party representative to act on their behalf in exercising their data rights.[83] This has inspired and enabled several organizations to provide services that help individuals exercise their DSARs, by acting on their behalf. Examples include the "Permissions Slip" mobile app by Consumer Reports.[84] The workflow is only partially automated (e.g., through the use of request templates) and still largely relies on human representatives contacting the companies on behalf of the user.[85]

It remains challenging for individuals, or their representatives, to exercise DSARs. One challenge is that it is cumbersome to submit these requests to multiple companies at scale.[86] To scale DSARs for the future, it must be possible to exercise these rights in a programmatic manner. For example, the Data Rights Protocol aims to develop a web standard to enable individuals to exercise DSARs provided under regulations such as the GDPR and the CCPA in an automated and programmatic manner.[87] Unlike Global Privacy Control ("GPC"), which is essentially a binary flag indicating an individual's "Do Not Sell" preference,[88] this proposal allows consumers to express fine-grained DSARs.[89] Other efforts include Advanced Data Protection Control ("ADPC"), which aims to serve as an alternative to cookie banners.[90] Similar to the Data Rights Protocol, ADPC could allow finer-grained privacy control than GPC.[91]

Another challenge involves company responses to the exercise of data subject rights. Companies may delay responding, provide inaccurate or incomplete information, or fail to respond altogether.[92] There is

---

83. CAL. CIV. CODE § 1798.185(a)(7) (West 2023).

84. PERMISSION SLIP, https://permissionslipcr.com [https://perma.cc/TKE2-Y6FG].

85. *See* Pegah Moradi, *An Early Look at How Companies Handle CCPA Requests Submitted by Authorized Agents*, CONSUMER REPS. (Aug. 22, 2022) https://innovation.consumerreports.org/an-early-look-at-how-companies-handle-ccpa-requests-submitted-by-authorized-agents [https://perma.cc/TH5Y-YHNR].

86. Daniel Solove, *The Limitations of Privacy Rights*, 98 NOTRE DAME L. REV. 975, 985 (2023) ("In many cases, an individual must exercise not just one right but several rights. These multiple rights must be exercised with hundreds if not thousands of organizations.").

87. Dazza Greenwood, Ryan Rix, Kevin Riggle, John Szinger & Ginny Fahs, *Data Rights Protocol*, GITHUB: CONSUMER REPS. INNOVATION LAB, https://github.com/consumer-reports-digital-lab/data-rights-protocol [https://perma.cc/2G89-XXRA].

88. GLOBAL PRIV. CONTROL, https://globalprivacycontrol.org [https://perma.cc/LR53-9H9X].

89. *Approach*, DATA RTS. PROTOCOL, https://datarightsprotocol.org/approach [https://perma.cc/A7FD-HNNN].

90. ADVANCED DATA PROT. CONTROL, https://www.dataprotectioncontrol.org [https://perma.cc/K867-G2EJ].

91. *Id.*

92. *See, e.g.*, *Action Taken Against SEVEN Organisations who Failed in Their Duty to Respond to Information Access Requests*, INFO. COMM'R'S OFF., https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/09/action-taken-against-seven-organisations-who-failed-in-their-duty-to-respond-to-information-access-requests [https://perma.cc/7W9B-D5CH].

currently no user application programming interface ("API") to verify whether companies implement the request (i.e., have fully disclosed, deleted, or corrected all maintained user data).

3. Privacy Representations "In the Wild"

Privacy representations also exist outside the well-defined world of privacy policies and DSARs. For example, companies make representations in press releases, blogs, and responses to journalists, as well as in privacy white papers and product documentation.[93] Furthermore, companies often disclose data collection in quarterly Form 10-K reports, as it pertains to company valuations.[94] Companies and their employees also are sometimes deposed or make representations in courts of law, adding representations to the public record via court documents.[95] Finally, there may be additional disclosures of data practices via the user interfaces of products (e.g., in a mobile OS permission dialog, where an app declares *why* it is asking for permission to access GPS location).[96]

While each of these forms of privacy representations can reveal important information, there is currently no systematic way to gather such data comprehensively and at scale. Future research, such as on natural language processing algorithms, could potentially make headway.

B. Measuring the Actual Behavior of a Company's Systems

The way a company promises to process data and the way it actually does so do not always align in practice. Researchers, including ourselves, have developed ways to discover how organizations' systems actually process data, through measurement.

1. Methodologies

Key challenges for regulating data privacy include the numerous modalities and ways in which data is collected, shared, and used by online systems, making it difficult to apply one auditing approach to all

93. *See, e.g.*, *Data Practices*, GOOGLE, https://safety.google/privacy/data/ [https://perma.cc/YW6V-K4WE]; *Amazon Sidewalk Privacy and Security Whitepaper*, AMAZON, https://www.amazon.com/gp/help/customer/display.html?nodeId=GRGWE27XHZPRPBGX [https://perma.cc/JA6Z-4UVT].

94. *See, e.g.*, Vizio Holding Corp., Annual Report (Form 10-K) 9 (Mar. 10, 2022).

95. *See, e.g.*, Letter from Google Legal to Österreichische Datenschutzbehörde (Apr. 9, 2021), https://noyb.eu/sites/default/files/2021-05/2021-04-09_Response_to_Austrian_DPA_-_NOYB_Complaints_b.pdf [https://perma.cc/QP8N-KLF9].

96. *See, e.g.*, *About Privacy and Location Services in iOS, iPadOS, and watchOS*, APPLE, https://support.apple.com/en-us/HT203033 [https://perma.cc/45C3-AL9H].

scenarios. There is no one-size-fits-all auditing methodology for all questions of interest (e.g., Does company X collect my personal data? Which data? How does it use this data? For what purpose?) or under all constraints (e.g., Is the audited system a black box or can we have access to its internals?). Nevertheless, various methodologies have emerged in the research community and are being applied across multiple ecosystems. This Section provides a summary of these methods and how our community has applied them in various auditing contexts. The principles and methods we describe below lend themselves to integration into regulatory frameworks and auditing implementations.

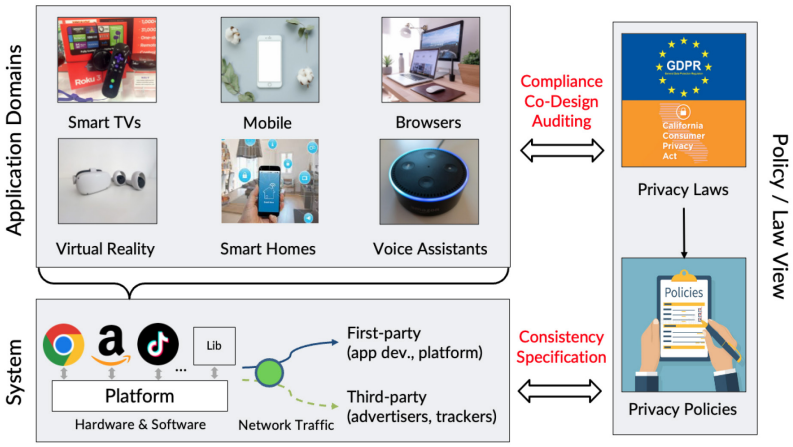

Figure 1: Overview of (1) auditing data collection practices, of platforms and applications, using network traffic monitoring at the edge of the network; and (2) checking the consistency with the corresponding privacy policies as well as the privacy law requirements.

The above figure depicts various end systems (e.g., smart TVs, mobile devices, browsers, VR devices, IoT devices) and their respective apps with which consumers typically interact. In all these ecosystems, personal data is (1) collected by software (e.g., operating systems, apps, analytics and other third-party libraries) running on these end-systems; (2) then sent over the Internet to first- and third-party servers for functional, advertising and tracking services, and many other purposes; and (3) further shared with other entities (e.g., cloud providers, data brokers) for personalization, advertising, and other monetization purposes. In this complicated and opaque tracking ecosystem, the only data collection and data flows that independent researchers can directly observe are at or close to the end systems themselves (e.g., in situ devices, apps, and network traffic they generate). Since we lack visibility into what

happens at companies' servers and how they share and process data, understanding this behavior requires measurement and inference approaches that observe and probe these systems from the edge of the network (referred to as "the edge" going forward).

Our team, as well as other researchers, has developed such approaches of measuring a system's data collection and use practices from the edge by controlling actions on the device and observing information flow in and out of the device. Specifically, the research community has followed three broad types of approaches for auditing data collection practices at the edge: (a) Direct Measurement of Data Collection, (b) Indirect Inference of Data Use, and (c) Company-Aided Measurement.

a. Direct Measurement

A large body of work obtains and analyzes the actual information flow observed out of an app, device, or platform using a range of techniques, including:

(1) **Static and dynamic code analysis**: In this approach, one conducts analysis on the computer code that runs on the device, e.g., an app or device software. In static analysis,[97] one analyzes the code to understand all the things that it *could* do if it were run on its intended device. A key challenge for this approach is that there is an enormous combination of inputs that consumers might provide to a device, and it is generally infeasible to explore how every input might lead to different software behavior. Furthermore, there is a gap between the set of all things that software *could* do compared to what software actually does when a consumer interacts with it.

Dynamic analysis[98] takes a complementary perspective and analyzes what software does when it actually runs on the intended system. For example, dynamic analysis could entail running an app on a smartphone and observing its network traffic. In this case, we learn what software actually does when a consumer (the researcher) interacts with it; however, we cannot explore all possible interactions that consumers can have with the software.

97. *See, e.g.*, William Enck, Damien Octeau, Patrick McDaniel & Swarat Chaudhuri, *A Study of Android Application Security*, 20 PROC. USENIX SEC. SYMP. 315, 315 (2011).

98. *See, e.g.*, William Enck, Peter Gilbert, Byung-Gon Chung, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel et al., *TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones*, 9 PROC. USENIX CONF. ON OPERATING SYS. DESIGN & IMPLEMENTATION 393, 393 (2010).

When analyzing software behavior today, researchers often apply both strategies to the extent possible.[99] Static code analysis informs us of what software could do, providing insights that guide our dynamic analysis to understand what the software actually does in practice. Our community has developed extensive automated analysis tools to facilitate this approach.[100] One key challenge is that, in many closed systems, the computer code is either obfuscated (i.e., scrambled in a way that makes it difficult to analyze statically[101] or unavailable (e.g., voice assistants prevent direct access to device software). In these cases, we cannot apply static analysis and are left with only dynamic analysis. By forcing such software to be open to analysis by qualified researchers and auditors, future regulation can fill in existing gaps.

(2) **Network traffic analysis**: Core privacy issues are implicated by the collection and sharing of data from end devices. Since, by definition, data leaves the device through the network interface, network traffic analysis is one of the primary tools to find "smoking guns" related to personal data being exposed to various parties (e.g., over the Internet or over local wireless communication).

Generally speaking, the key challenges here are (1) getting access to network traffic and (2) being able to interpret whether there is personal data in such traffic. Our team and community have built upon existing tools and extended others to enable collection of network traffic across various modalities that include web browsers,[102] mobile devices,[103]

---

99. *See, e.g.*, Umar Iqbal, Steven Englehardt & Zubair Shafiq, *Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors*, 2021 IEEE SYMP. ON SEC. & PRIV. 1143, 1143; Valentino Rizzo, Stefano Traverso & Marco Mellia, *Unveiling Web Fingerprinting in the Wild Via Code Mining and Machine Learning*, 2021 PROC. ON PRIV. ENHANCING TECHS. 43, 43.

100. *See, e.g.*, Steven Englehardt & Arvind Narayanan, *Online Tracking: A 1-Million-Site Measurement and Analysis*, 2016 PROC. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 1338, 1338; Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian & Zubair Shafiq, *AdGraph: A Graph-Based Approach to Ad and Tracker Blocking*, 2020 IEEE SYMP. ON SEC. & PRIV. 763, 763.

101. *See, e.g.*, Philippe Skolka, Cristian-Alexandru Staicu & Michael Pradel, *Anything to Hide? Studying Minified and Obfuscated Code in the Web*, 2019 WORLD WIDE WEB CONF. 1735, 1735.

102. *See, e.g.*, Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq & Carmela Troncoso, *WebGraph: Capturing Advertising and Tracking Information Flows for Robust Blocking*, 31 PROC. USENIX SEC. SYMP. 2875, 2879 (2022); Iqbal et al., *supra* note 99, at 768.

103. *See, e.g.*, Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich et al., *Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem*, NETWORK & DISTRIB. SYS. SEC. SYMP., Feb. 2018, at 1; Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout &

smart TVs,[104] smart speakers,[105] VR headsets,[106] and other IoT devices. Combined with the dynamic code analysis approaches mentioned above, we developed techniques to automatically interact with hardware and software and capture all Internet or other wireless traffic exchanged between these devices and others.[107] This typically involves instrumenting routers or building custom network-traffic collection devices and software for a testbed environment (e.g., the Mon(IoT)r Testbed[108]).

The second challenge is how to determine whether personal data is being transmitted in encrypted network traffic. End-to-end encryption rightfully protects data from eavesdroppers (e.g., ISPs or other users of public WiFi access points) by ensuring that only the endpoints (software on the client or server) can correctly interpret the data in the connection.[109] However, they also prevent researchers and regulators from seeing the content of communications, making audits difficult.[110] A common approach to addressing this problem is to modify the targeted software to allow for decryption of network traffic at the observation points (e.g., at the wireless

---

David Choffnes, *ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic*, 14 PROC. INT'L CONF. ON MOBILE SYS., APPLICATIONS & SERVS. 361 (2016); Anastasia Shuba, Anh Le, Emmanouil Alimpertis, Minas Gjoka & Athina Markopoulou, *AntMonitor: A System for On-Device Mobile Network Monitoring and Its Applications*, PROC. 2015 ACM SIGCOMM WORKSHOP ON CROWDSOURCING & CROWDSHARING BIG (INTERNET) DATA, Aug. 2015, at 15 [hereinafter *AntMonitor*]; Anastasia Shuba & Athina Markopoulou, *NoMoATS: Towards Automatic Detection of Mobile Tracking*, 2020 PROC. ON PRIV. ENHANCING TECHS. 45; Anastasia Shuba, Athina Markopoulou & Zubair Shafiq, *NoMoAds: Effective and Efficient Cross-App Mobile Ad-Blocking*, 2018 PROC. ON PRIV. ENHANCING TECHS. 125.

104. *See, e.g.*, Hooman Mohajeri Moghaddam, Gunes Acar, Ben Burgess, Arunesh Mathur, Danny Yuxing Huang, Nick Feamster et al., *Watching You Watch: The Tracking Ecosystem of Over-the-Top TV Streaming Devices*, 2019 PROC. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 131; Janus Varmarken, Hieu Le, Anastasia Shuba, Zubair Shafiq & Athina Markopoulou, *The TV is Smart and Full of Trackers: Measuring Smart TV Advertising and Tracking*, 2020 PROC. ON PRIV. ENHANCING TECHS. 129.

105. *See, e.g.*, *Echos*, *supra* note 1.

106. *See, e.g.*, Trimananda, *supra* note 77.

107. Tianrui Hu, Daniel Dubois & David Choffnes, *BehavIoT: Measuring Smart Home IoT Behavior Using Network-Inferred Behavior Models*, 2023 PROC. ACM INTERNET MEASUREMENT CONF. 421, 421.

108. *Mon(IoT)r Testbed*, MON(IOT)R RSCH. GRP., https://moniotrlab.khoury.northeastern.edu/tools/ [https://perma.cc/QGK3-EYXX].

109. *What is Encryption?*, GOOGLE CLOUD, https://cloud.google.com/learn/what-is-encryption [https://perma.cc/U6RX-59QK].

110. Daniel J. Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes & Hamed Haddadi, *When Speakers Are All Ears: Characterizing Misactivations of IoT Smart Speakers*, 2020 PROC. ON PRIV. ENHANCING TECHS. 255, 261.

> router), an approach called MITM.[111] While this approach has been used for desktops, laptops, and mobile devices, it is not always feasible to make these changes on closed systems (e.g., on smart speakers). Furthermore, even with access to the decrypted contents of network traffic, companies can encode or hide the data being collected.[112] Our team and others have used a variety of approaches to address the challenge of finding personal data in network traffic, including doing analysis using multiple encodings, using search terms that identify where personal data is located, and isolating identifiers used as "pseudonymous" substitutes for personal data.[113]

By using one or more of these techniques, the auditor can directly observe the data collection and sharing practices of the studied hardware and software, revealing what data types are collected, to what destination they are sent, whether these destinations belong to the app or platform company or to third parties such as advertisers and trackers, and other data practices. Our team has employed these techniques and published various revealing findings about data tracking practices, which are highlighted in the next Part.

b. Indirect Inference

Sometimes it may not be possible to directly observe what is collected by an end device (e.g., a laptop, smartphone, or IoT device). It may be difficult to intercept or decrypt the traffic,[114] or information about how data is used may not be available to independent parties because it is done at servers beyond the reach of analysts.[115] Other times, we are more interested in how companies share (the same or derived) information with other parties (e.g., data brokers). In both cases, the behavior of interest is not directly observable from the edge and can only be inferred.[116]

Researchers have developed methodologies to infer, from the edge of the network, whether data is collected and whether it is used by first

---

111. Short for "monster in the middle," a reference to the fact that a "middlebox" (i.e., something that is not an endpoint) is intercepting traffic that is intended to be viewed only by endpoints. *Monsters in the Middleboxes: Introducing Two New Tools for Detecting HTTPS Interception*, CLOUDFLARE, https://blog.cloudflare.com/monsters-in-the-middleboxes/ [https://perma.cc/BZ43-9K5G].

112. Ren et al., *supra* note 56.

113. *See supra* note 103 and accompanying text.

114. *See, e.g.*, Amogh Pradeep, Muhammad Talha Paracha, Protick Bhowmick, Ali Davanian, Abbas Razaghpanah, Taejoong Chung et al., *A Comparative Analysis of Certificate Pinning in Android & iOS*, 2022 PROC. ACM INTERNET MEASUREMENT CONF. 605, 605.

115. Cook et al., *supra* note 50, at 66.

116. *Echos*, *supra* note 1, at 572.

or third parties.[117] The general principle is relatively simple: if our hypothesis is that personal data is being used by an entity for some purpose, we seek out evidence that it is. In the case of ad personalization, we can indirectly infer use of personal data when we see ads that are targeted based on that personal data. Likewise, we can infer sharing of data when we see such targeted ads from parties that never directly received personal data from an individual.

While the principle of indirect inference is simple, the correct implementation requires carefully designed and executed experiments that can reveal such data usage or sharing with high confidence. The general approach to address this is the following: the auditor runs software on the device under investigation and (1) controls user actions on the system, (2) observes responses from the first party and/or related systems, and (3) analyzes the results to infer whether data is collected, used, and/or shared for particular services. For (1), researchers typically control user actions by interacting with a service and exposing personal data and interests to it. We refer to any data that the service learns and stores about this user as a profile of the user. We generally want to see how services store information about users with different interests, so we repeat this process with multiple "fake" users (often called "sock puppets"[118]) and generate multiple user profiles accordingly. For (2), we might observe whether different profiles receive different personalized services or advertising (particularly compared to a profile where no data was shared); the latter could indicate data use via personalization. For (3), we can infer which entities provided personalized content and whether the data used by those entities was gathered directly from users or obtained from another entity, indicating data sharing.

A common thread in the scientific literature is the need to deal with noise: differences in ads or other information provided to different users that may not be due to personalization. For instance, ads are often displayed as a result of real-time auctions whose outcomes are not necessarily consistent from one auction to the next.[119] Such differences may occur when loading ads even for the same user on the same webpage. Our team and others in the community have developed a suite of methods to deal with such noise, using a combination of large

117. Jingjing Ren, Martina Lindorfer, Daniel J. Dubois, Ashwin Rao, David Choffnes & Narseo Vallina-Rodriguez, *Bug Fixes, Improvements, . . . and Privacy Leaks: A Longitudinal Study of PII Leaks Across Android App Versions*, NETWORK & DISTRIB. SYS. SEC. SYMP., Feb. 2018, at 1; Ren et al., *supra* note 56.

118. Christian Sandvig, Kevin Hamilton, Karrie Karahalios & Cedric Langbort, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms* 13 (unpublished manuscript).

119. Cook et al., *supra* note 115, at 66.

numbers of controlled experiments and statistical analysis to limit the potential impact of noise on our inferences.[120]

c. Company-Aided Measurement

In the direct and indirect measurements above, the system under investigation is a black box into which the auditor has no privileged access. However, via both compulsion (e.g., consent orders) and voluntary action (e.g., for public relations), companies may assist with the process of auditing their systems for compliance with representations and regulation. This ranges from extreme transparency, where the company publishes its algorithm and open sources its code, to the more typical choice of providing APIs for measurements or special access to hardware or software for researchers and auditors. For example, Facebook had provided a "white-hat" access to their mobile app for researchers, which allowed them to decrypt the network traffic and look for privacy and security issues.[121] In an audit, Pymetrics, the talent assessment and hiring platform using AI, gave researchers access to source code and company data to evaluate fairness claims made by the company.[122] Similarly, the Apple Security Research Device Program gives qualified researchers access to "specially fused iPhones" to help identify iOS security vulnerabilities.[123]

An important aspect of company-aided measurement is ensuring that the results from such measurements match those seen "in the wild." Even if a company provides transparency into a portion of their software or hardware systems, this does not necessarily mean that observed behavior matches that seen by consumers who potentially interact with different systems. Here, a "trust but verify" approach serves the auditor well. One can use insights gleaned from increased transparency to better understand expected system behavior, then use black-box approaches to understand whether the corresponding off-the-shelf hardware/software behavior matches.

---

120. *See, e.g.*, *Echos*, supra note 1; Zengrui Liu, Umar Iqbal & Nitesh Saxena, *Opted Out, Yet Tracked: Are Regulations Enough to Protect Your Privacy?*, 2024 PRIV. ENHANCING TECHS. 280, https://petsymposium.org/popets/2024/popets-2024-0016.pdf [https://perma.cc/4VWF-X9T8].

121. Mohit Kumar, *New Settings Let Hackers Easily Pentest Facebook, Instagram Mobile Apps*, HACKER NEWS (Mar. 26, 2019), https://thehackernews.com/2019/03/facebook-whitehat-setting-hackers.html [https://perma.cc/WD3G-XZYT].

122. Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, 2021 PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 666, 671.

123. *Apple Security Research Device Program*, APPLE, https://security.apple.com/research-device [https://perma.cc/9EJD-XQJD].

IV. USING THE SCIENTIFIC RESULTS FOR ACCOUNTABILITY

The results of a scientific approach to surfacing privacy representations and clarifying the operation of systems can be used to hold tech companies accountable. We place all our efforts into two broad categories — taking advantage of existing policy levers and providing justifications for new rules.

A. Existing Uses of Research

Broadly speaking, once researchers have found evidence that companies are lying, acting dangerously, or violating the Fair Information Practices, they have at least three options that can coexist and overlap: they can (1) disclose their findings, (2) work with government and industry, and/or (3) help the public take action themselves.

First, when considering actions to take after discovering harmful behavior, researchers must incorporate responsible disclosure principles. At the heart of such principles is a key question: at what point would public disclosure lead to more risk for affected consumers than a private disclosure to the responsible party to first mitigate the problem? For many privacy issues that are also security concerns (e.g., consumer passwords exposed in plaintext network traffic where eavesdroppers on public Wi-Fi can see them), we recommend following a responsible disclosure approach that pairs an initial private disclosure with an eventual public one after a remediation window. For example, we disclosed password exposure vulnerabilities privately to affected companies with a deadline (sixty–ninety days, with extensions granted with reasonable justification)[124] for remediation, after which we went public. The idea was that private disclosure and remediation prevented additional harm from attackers knowing that there was a vulnerability to exploit. However, if a company was unresponsive, we reasoned that no remediation was forthcoming, so the better action for consumers was to go public to ensure awareness (and to encourage them to avoid using the affected software). For cases where privacy issues are not immediate security vulnerabilities (e.g., collection of device identifiers without consent, as opposed to exposing consumer passwords in unencrypted network traffic), the responsible disclosure calculus is different. Here, public disclosure before remediation causes no additional harm to consumers. In such cases, we have conducted public outreach by publishing articles in peer-reviewed scientific venues[125] and in the popular press, disclosing issues to vendors, developing

124. For example, one company needed extra time to reach out to physicians to get them to update their software before public disclosure, so we granted this extension. Ren et al., *supra* note 103, at 11.

125. *See, e.g.,* Ren et al., *supra* note 103; Dubois et al., *supra* note 110.

tools to help users avoid privacy harms unilaterally, and informing stakeholders in the public and private sectors — in some cases simultaneously.

While responsible disclosure principles from the computer security community are useful guideposts for researcher actions after discovering privacy issues, we can do more to leverage hard-won insights from computer security. For example, one way we envision researchers could disclose the results of their research is through a national privacy vulnerability database ("NPVD"), akin to the national vulnerability database ("NVD").[126] The database could include details on the nature and severity of the privacy issue, its impact on user privacy, the affected software or hardware and the corresponding organization (along with the links to each organization's respective privacy policy), and any mitigations or workarounds that consumers can use to protect themselves. Like NVD, NPVD could also include a severity rating system that would help users understand the seriousness of each privacy vulnerability.[127] The maintenance of the national privacy vulnerability database could be the responsibility of a government agency (e.g., FTC's Office of Technology Research and Investigation)[128] or a third-party organization. This organization would be responsible for collecting, analyzing, and publishing data related to privacy vulnerabilities from different stakeholders in a transparent manner. Clear guidelines and governance structures should be implemented to ensure that the database is maintained in a fair and unbiased manner. Related to the NPVD, we also envision the adoption of privacy bug bounty programs, akin to the increasingly popular (security) bug bounty programs run by many organizations.[129] There has been an extensive discussion in the community about how to run such programs successfully, including clear specifications of scope.[130] Regulators can help here as well, since they can help forge uniform requirements around privacy bug bounty programs to ensure their fair and effective use.

---

126. *National Vulnerability Database*, NAT. INST. STANDARDS & TECH., https://nvd.nist.gov [https://perma.cc/B5G4-TKRQ].

127. *Vulnerability Metrics*, NAT. INST. STANDARDS & TECH., https://nvd.nist.gov/vuln-metrics/cvss [https://perma.cc/2MNY-ESHC].

128. *Office of Technology Research and Investigation*, FTC, https://www.ftc.gov/about-ftc/bureaus-offices/bureau-consumer-protection/our-divisions/office-technology-research-investigation [https://perma.cc/7HRU-NY94].

129. For example, BUGCROWD, https://www.bugcrowd.com [https://perma.cc/8PDY-NXMQ], and HACKERONE, https://www.hackerone.com [https://perma.cc/VYM8-JZME], provide a platform where security researchers can be rewarded for identifying and privately disclosing security problems for online systems run by corporations.

130. *See Good Policies*, HACKERONE, https://docs.hackerone.com/organizations/good-policies.html [https://perma.cc/3AMY-YRES]; *Safe Harbor FAQ*, HACKERONE, https://docs.hackerone.com/organizations/safe-harbor-faq.html [https://perma.cc/7G3B-N6GG].

The second way that researchers can use their findings for tech accountability is to work with lawmakers, regulators, standards bodies, auditors, and developers to help ensure compliance with privacy rules. For example, we routinely engage with the privacy-enhancing browser extension software developer community to share our findings with regard to the limitations of the maintenance processes of crowdsourced filter lists.[131] We have also engaged with the browser standards community (e.g., W3C WebExtensions Community Group[132] and Privacy Interest Group[133]) to appraise the community of our research findings regarding ongoing standardization work. In addition to working in the browser ecosystem, we have worked with app developers and app store maintainers to address observed harms. For example, we responsibly disclosed cases where apps exposed consumer passwords in plaintext.[134] Additionally, we reported apps violating app store policies so they could be removed, and some of our findings have led to new app store policies (e.g., banning screen recording).[135]

We regularly speak with staff for lawmakers who are considering privacy legislation and give feedback on draft language in proposed legislation. We have engaged with regulators (in particular, the FTC) via regular participation in PrivacyCon events, conversations with FTC technologists to explain our findings in more detail, and public code and data sharing to support efforts to reproduce our findings. We also hosted our own workshop on the specific challenges of regulating privacy and security for IoT devices.[136] We have presented testimony in front of Congressional committees, submitted comments for proposed rulemaking, and served on government committees to present recommendations to lawmakers on topics like facial surveillance.[137]

Finally, researchers can use their findings to help people take action for themselves to keep tech companies accountable. One way to help is to provide tools for both the public and developers to use to

---

131. *See* AD-FILTERING DEV SUMMIT, https://adfilteringdevsummit.com [https://perma.cc/E64E-XV2N].

132. *WebExtensions Community Group*, WORLD WIDE WEB CONSORTIUM, https://www.w3.org/community/webextensions [https://perma.cc/7GLY-BXZP].

133. *Privacy Interest Group*, WORLD WIDE WEB CONSORTIUM, https://www.w3.org/Privacy/IG [https://perma.cc/8WPA-HT83].

134. *See* Ren et al., supra note 103.

135. *See* Elleen Pan, Jingjing Ren, Martina Lindorfer, Christo Wilson & David Choffnes, *Panoptispy: Characterizing Audio and Video Exfiltration from Android Applications*, 2018 PROC. PRIV. ENHANCING TECHS. 33, 44–45; *see also* Zach Whittaker, *Apple Tells App Developers to Disclose or Remove Screen Recording Code*, TECHCRUNCH (Feb. 7, 2019, 4:43 PM), https://techcrunch.com/2019/02/07/apple-glassbox-apps [https://perma.cc/NSZ4-T9XA].

136. *SPLICE and ProperData Host Joint IoT Workshop*, PROPERDATA (Apr. 13, 2022), https://properdata.eng.uci.edu/2022/04/13/splice-and-properdata-host-joint-iot-workshop/ [https://perma.cc/55H4-NZKS].

137. SPECIAL COMMISSION TO EVALUATE GOVERNMENT USE OF FACIAL RECOGNITION TECHNOLOGY IN THE COMMONWEALTH 3 (2022), https://archives.lib.state.ma.us/items/66158ea1-78f7-4a3e-80c0-9e2051dcb71b [https://perma.cc/NTG9-Y4HS].

understand software and data flows.[138] It can also mean collaborating on litigation or even helping people exercise their rights of transparency, accuracy, and deletion as data subjects. In our ReCon and AntMonitor work, we provided users with software that can run on their mobile devices and reveal personal data transfers and the entities that receive that data, along with the capability to block it.[139] In addition, we built websites that show consumers what data is collected by apps, how this collection changes over time, and how severe the data exposure is based on individual preferences.[140] We also regularly engage journalists to help spread the word, bringing our findings to a larger audience and encouraging consumers and lawmakers to take action to remediate observed harms.[141]

B. Justifying New Rules

Sometimes researchers will uncover misleading representations and dangerous actions that current privacy law fails to contemplate. Other times, researchers' findings demonstrate the limits of our current privacy rules and our limited ability to understand the scope of privacy issues and the role of technology in either exacerbating or remediating such issues. In these circumstances, research justifies new substantive and structural rules for better tech accountability.

Substantively, research of a significant problem can highlight the shortcomings of current rules and the need for new ones. For example, the notice requirements in the GDPR and the CCPA have proven insufficient. People are typically unable to process and comprehend the privacy policies, which prevents them from making informed choices about their use of services and applications.[142] Even as a transparency and accountability mechanism, privacy policies often lack specificity over what personal information is collected and how, leaving consumers uncertain about the related privacy risks.[143] Additionally, privacy policies often lack transparency about what categories of personal information are required to provide service or app functionality versus

---

138. For example, the ReCon project makes such data available at RECON, https://recon.meddle.mobi [https://perma.cc/JP6M-LQ2Y], and similar data was provided at *AppCensus: Learn the Privacy Costs of Free Apps*, INT'L COMPUT. SCI. INST., https://www.icsi.berkeley.edu/icsi/projects/privacy/app-census [https://perma.cc/T8VF-VUFM].

139. Ren et al., *supra* note 103, at 361; *AntMonitor*, *supra* note 103, at 15. These tools allowed users to view personal data being shared by mobile apps with other parties over the Internet, and to block such data flows by destination and type of data (e.g., geolocation or unique identifier).

140. *Should You Update Your App?*, RECON, https://recon.meddle.mobi/appversions/tool.html [https://perma.cc/X4RY-6Y45].

141. *See* HARVEST (Indevu Films 2017).

142. *Proposal*, *supra* note 45, at 280–89.

143. *Id.*

non-functional purposes such as advertising, frustrating consumers' attempts to balance functionality and privacy.[144] Privacy policies often fail to disclose sufficient information about the sharing of personal information, impeding consumers' ability to understand the degree of identifiability of their shared information, to determine the associated privacy risks, or to follow the dissemination of their personal information throughout the data ecosystem.[145]

A comprehensive consumer privacy law should remedy these shortcomings of the GDPR and the CCPA. One approach would require disclosure of the purposes for collecting and sharing each category of personal information.[146] Alternatively, perhaps the entire purpose of disclosure and consent should be revisited and replaced with a substantive duty of loyalty that would prioritize people's best interests and compel more transparency to regulators, more forthrightness to people, and less room for bad actors seeking to justify dubious business practices, data flows, and design strategies.[147]

This approach and our findings generally highlight a greater need for technical support to interrogate Internet-connected products and their network traffic. For example, lawmakers could support requirements that compel vendors to provide standard ways for qualified researchers and regulators to access product hardware, software, and/or data transmitted over the Internet. For example, when vendors provide this functionality on products made available only for such analysis, independent parties can pursue rigorous verification of consumer protections without breaking privacy and security protections for products placed in the hands of consumers. This functionality already voluntarily exists to some degree (e.g., the Apple Security Research Device Program[148]), and expanding such functionality to more products will further improve security and privacy for consumers.

Privacy policies often use non-standardized definitions of personal information that do not align with those in the GDPR or the CCPA or even with each other, leaving consumers confused about what constitutes personal information.[149] Privacy policies often include assertions about the anonymity of personal information that exceed both the technical abilities and legal definitions of anonymization and of de-

144. *Id.*

145. *Id.*

146. *Id.*

147. *See generally* Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961, 961 (2021); Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 EMORY L.J. 985, 985 (2022); Woodrow Hartzog & Neil Richards, *Legislating Data Loyalty*, 97 NOTRE DAME L. REV. REFLECTION 356 (2022); Woodrow Hartzog & Neil Richards, *Privacy's Constitutional Moment and the Limits of Data Protection*, 61 B.C. L. REV. 1687, 1696 (2020); Paul Ohm, *Forthright Code*, 56 HOUS. L. REV. 471 (2018).

148. *Apple Security Research Device Program*, *supra* note 123.

149. *Proposal*, *supra* note 45, at 263–67.

identification.[150] A comprehensive consumer privacy law should remedy these shortcomings of the GDPR and the CCPA by defining not only personal information and de-identified information, but also pseudonymous information and nontrackable information.[151] It should require disclosure of the form of personal information used and shared[152] and properly incentivize the use of such forms of information over the use of reasonably identifiable information.[153]

Statutory definitions of personal information, de-identified information, pseudonymous information, and nontrackable information should reflect the findings in the computer science literature regarding the identifiability of different forms of information.[154] Some attempts have been made to bridge computer science concepts with legal definitions and establish interdisciplinary meanings.[155] However, privacy statutes have failed to incorporate these findings and have instead relied on the oversimplified and imprecise categorization of information based solely on whether it is reasonably linkable.

V. CONCLUSION

In this Essay, we have proposed a scientific approach for academic researchers to help regulators and consumers keep the ever-evolving tech industry accountable for their privacy practices. This approach involves surfacing a company's privacy representations, measuring the actual behavior of a company's systems, and using these scientific results for greater accountability. By working with academic researchers and public interest technologists, regulators and consumers can better enforce privacy rules. Our proposed collaboration between researchers, regulators, and consumers could lead to more robust and effective tech regulation, benefiting both consumers and the tech industry itself.

150. *Id.*
151. *Id.*
152. *Id.* at 280–89.
153. *Id.* at 270–73.
154. Scott Jordan, *Aligning Legal Definitions of Personal Information with the Computer Science of Identifiability*, PROC. RSCH. CONF. ON COMMC'N, INFO. & INTERNET POL'Y, Sept. 2021, at 1.
155. *See* Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification*, 56 SANTA CLARA L. REV. 593, 594 (2016); *see also* SIMSON L. GARFINKEL, NAT'L INST. STANDARDS & TECH., INTERNAL REP. No. 8053, DE-IDENTIFICATION OF PERSONAL INFORMATION 1–2 (2015).