A Standardized Model for Transparent Data Privacy

1st Brian Tang
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA
bjaytang@umich.edu

***Abstract*—Data privacy has been a major concern at the forefront of numerous discussions on big data and machine learning. Many companies gather vast amounts of user data (e.g., location data, browsing behavior, consumer purchases, etc.) and create machine learning (ML) models using this data. Unfortunately, the current "notice and consent" model of data privacy raises serious issues with transparency and security. For example, data practices are hidden within long and difficult-to-understand privacy policies; explanations regarding the purposes of data collection can often be vague or inaccurate. As a result, users are often unaware of risky and unethical data practices until data breaches and data misuses appear in the media.**

Two key questions regarding data privacy remain unaddressed: 1) *where does my data go* and 2) *what is it being used for*? While improvements in legislature such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) have set compliance guidelines for these issues, privacy policies in their current state are still a fundamentally broken form of informing users. We propose that privacy can be made more transparent by regulating privacy polices to use an easily parse-able standardized privacy disclosure. This would not only drastically simplify the challenges associated with interpreting and understanding natural language, it would enable the development of privacy tools to easily interpret, summarize, and display information to users. Furthermore, it would allow users to finally understand the interactions between the various entities that collect and aggregate their data.

In this paper, we provide three main arguments: 1) Privacy policies should be replaced with a standardized form. Data processors should disclose every data transaction (or at least each processed data type) rather than providing vague explanations of where user data goes. 2) Easily understandable summaries of privacy can easily be generated by aggregating the transactions in this standardized form. 3) The actual transmissions of data should be automatically regulated and manually verified to save time and money. We discuss the rationale behind this overhaul to the current state of privacy policies.

I. Introduction

Data privacy has been a major concern at the forefront of many discussions on cutting-edge technology. Many companies have gathered vast quantities of user data (e.g., location data, browsing behavior, consumer purchases, etc.) and create models using this data [1]. Unfortunately for users, the data collection and usage practices of companies impose significant issues with transparency and security. For example, data practices may be hidden among long and unreadable privacy policies; explanations regarding the usage or purposes of data collection can often be vague or inaccurate [2]. In other scenarios, such as providing consent or opting-out of data collection, users are often nudged into selecting choices harmful to their privacy in what are known as dark patterns [3, 4]. As a result, users are often unaware of risky data practices or misuses until news about major data breaches or data misuses appear [5, 6].

There are major indicators that point towards the ability for online data privacy laws to be exploited or loop-holed [7, 8]. Until users are properly informed and not misled by exploiting their cognitive biases [9], they will continue to act as irrational agents [10]. Thus, to protect users and reinforce users' trust in data collectors, there needs to be a greater level of transparency and accountability. Taking inspiration from social norms, privacy norms, and regulatory administrations from other industries, we can create a standardized model of presenting privacy information to users. This can allow users to easily understand how their data will be collected and used, easily opt-out of data collection, and request deletion of their data. Additionally, regulators should be able to automatically monitor and arbitrate any privacy violations detected or reported by users.

Existing privacy laws such as the General Data Protection Regulation (GDPR) have been created to protect data privacy. While there has been some improvement [11], regulators are struggling to enforce these protections on a large-scale [12–14]. Researchers have also developed solutions to read and interpret privacy policies and automate consent mechanisms [15, 16]. Additionally, several designs have been suggested for privacy interfaces for Internet of Things (IoT) devices [17]. However, all of these solutions address specific problems and have not yet seen widespread user or regulatory adoption. Furthermore, without a way for regulators to analyze and enforce privacy violations, data collectors may opt to omit or falsify information without fear of penalties.

We will present our arguments and solutions by (1) formally identify the biggest issues surrounding data privacy, (2) motivating a standardized design and model of privacy by identifying the shortcomings of existing solutions, (3) modeling the decision-making processes of users/regulators/companies as agents, (4) providing a design standard which all data collectors must follow, and (5) providing solutions for the adoption, enforcement, and personalization of our privacy design standard. Throughout this paper, we will discuss potential shortcomings and limitations to each of the provided arguments.

The data transactions model will be characterized by the main entities involved in data collection and usage. These

include data owners, data collectors, privacy regulators, and other stakeholders. This model has its foundations rooted in the contextual integrity principle presented by Nissenbaum et al. [18].

The design standard will attempt to provide simple and intuitive representations of each of these data flows and usage purposes. The system should be designed to be easy for users to interpret, easy for data collectors to provide information about, and automated for regulation. It will present only the most concerning potential privacy hazards to users and provide a comprehensive overview to regulators.

In this paper, we argue that online data privacy should be 1) formally modeled, automatically regulated, and presented to users in easily understandable formats. While recent laws and regulations have helped better protect user privacy, these laws are still insufficient in regards to transparency, enforcement, and scalability.

1) Privacy policies should be replaced with a standardized form. Data processors should disclose every data transaction (or at least each processed data type) rather than providing vague explanations of where user data goes.
2) Easily understandable summaries of privacy can easily be generated by aggregating the transactions in this standardized form.
3) The actual transmissions of data should be automatically regulated and manually verified to save time and money. We discuss the rationale behind this overhaul to the current state of privacy policies.

II. Background

A. Brief Primer in the History of Privacy

The right to privacy is rooted in the constitution underlying American democracy in its First, Third, Fourth, and Fifth amendments. Over the course of the 20th century, many laws protecting the individual's right to privacy related to mail, educational records, health records, and more have been passed (FERPA, COPPA, etc.). Privacy limits government authority while preserving the individual's ability to practice freedom of thought and speech. On a personal level, privacy allows people to develop personal relationships and cultivate trust without fear of being eavesdropped. Invasion of privacy occurs whenever data is gathered on a person who has a reasonable expectation of privacy. For example, law enforcement is required to obtain a warrant when searching personal property for evidence related to a crime.

With the advent of the Internet, 2.5 exabytes of data is being created every day (2.5 quintillion bytes) [19]. Online services have adopted the current paradigm of Notice and Consent [20], where users are informed and asked to accept terms regarding the processing of their data. However, surveys estimate that only 91% of consumers skip reading legal terms and conditions [21]. Further, it would take a person 76 work days to actually read through all the privacy policies they encounter over the course of a year [22]. These results indicate that online settings lack a reasonable expectation of privacy. Rather, consumers place their trust in the companies running online services to protect and not misuse their data.

B. Surveillance Capitalism

Surveillance capitalism is an economic model which asserts that personal information is scraped and packaged to sell to others [23]. It is a unique business model that does not rely on establishing the traditional producer-consumer relationship. In this way, trust and reciprocity is no longer needed between users and service providers. Rather, users become sources of data extraction which are used to construct models of current and future market trends/behaviors. This product is then sold to other enterprises in order to train machine learning models, personalize and target advertisements, aggregate data, and create surveillance networks. Overtime, the granularity of collected data has only gotten more complex and detailed. Small behavioral tendencies (e.g., scrolling behavior, clicking behavior, punctuation use, emotions) have been collected to fine-tune recommendation and personalization models. Recent advances in machine learning (ML), particularly in deep neural networks (DNNs), have allowed the creation of these precise classification, behavioral, prediction, and recommendation models.

As a result of the economic successes of surveillance captialism, many companies built around collecting, aggregating, and selling data about users have emerged (e.g., Palantir, Comscore, Alteryx, Facebook, Google). This is problematic for the following reasons: 1) anonymity in advertising identifiers is lost, 2) users are uninformed about the different data vendors and customers, and 3) user data will be replicated across many platforms. The purpose of using advertising identifiers instead of personal identifiable information is to maintain the anonymity and privacy of users. However, when multiple sources of data is aggregated under a single advertising identifier, the original identity of a user can be reconstructed. For example, the simple combination of demographics like ZIP code, gender, and date of birth can uniquely identify most Americans [24]. Additionally, users are often unaware of the nuances in privacy policies, especially third-party sharing agreements. This can lead to confusion and feelings of betrayal upon realizing their data has been shared with untrustworthy organizations. Finally, the centralization and resharing of data can complicate the enforcement of "the right to be forgotten", or the deletion of data in online platforms.

C. Usable Privacy

Privacy policies and end user license agreements have been notorious for being long, vague, and difficult to read. Recent studies have found that most privacy policies require a college-level reading skills to effectively understand and interpret these policies [25, 26]. Evaluations demonstrate that policies have only grown in length and complexity after the enforcement of the GDPR [11]. As a result, researchers have explored design methodologies to improve the privacy and transparency of data processing systems. In particular, Spagnuelo et al. define 8 potential metrics for measuring transparency in systems

design [27]. These include: accuracy (Proportion of statements that are accurate/true), currentness (Amount of time between occurrence in system and information provided to the user), conciseness (Sentence length, total word count, total sentence count, semantics), detailing (Answering of questions such as: what? who? why? when? to whom? which?), readability (Pretty much same as conciseness but also syllable count), availability (Ease of actions user must perform, number of actions user must perform), portability (Available in any open format, available as structured data, available as non-proprietary format, uses URI, based on linked data), and effectiveness (Satisfiability of the mechanism's outputs and goals).

Another important question that needs to be addressed is the actual and perceived value of user data. A commonly studied issue is the "privacy paradox" [28] which demonstrates that although users value privacy, their online actions suggest that they only value data privacy at a few cents. More recent research has shown that users do not act rationally online due to bounded rationality and limited choice [9, 10, 29]. Users often make harmful privacy decisions as a result of not being properly informed or from resigning due to a lack of alternative options. Studies have investigated user awareness and their perceived valuation of data [30]. For example, their surveys have indicated that only 23-27% of people are aware that they are sharing social network friends lists, location, and web searches. In the United States, people would only pay greater than $50 to protect their government identification data. Everything else – purchase history, web search history, etc. is valued at under $50. The survey also gauged user trust in different types of organizations and their handling of data, putting doctors and financial firms at the highest level of trust, whereas social media and entertainment firms remained on the lowest level of trust.

Overall, it is difficult for users to know how much of their data is collected and how it is used. Users often have incomplete or asymmetric information (they are the worse informed party). Users find it impossible to assess what security or privacy vulnerabilities they might expose themselves to, and the privacy/security decisions they need to make a rife with trade-offs, complexity, and nuance. On top of this, security and privacy is rarely the end-users' primary goal. "Privacy by design" asserts that data controllers should be more ethical when handling data. This means more transparency, data minimization, anonymization, and reduction of data. From a user experience perspective, privacy can be enhanced by using "nudges". These nudges can use incentives, defaults, and feedback that guide users towards making privacy preserving decisions [31].

Unfortunately, the same design principles underlying "privacy by design" can be inverted to produce a set of "dark design strategies", where the goal is to maximize, publish, centralize, preserve, obscure, deny, violate, or fake data [4]. Users are susceptible to these "dark patterns" in user interface (UI) design. Dark patterns use commonly explored UI and HCI mechanisms such as nudging, persuasion, heuristics/biases, and cognitive dissonance to confuse the user and guide them towards making a decision harmful to their data. Some examples of these dark patterns include using confusing jargon/UIs, harmful defaults (opt-out), forced account registration, hidden legal stipulations, undeletable accounts, address book leeching (uploading contacts), and shadow user profiles (collecting information about unregistered people). The goal of these dark patterns is to ensure that users make irrational decisions online as a result of UI design taking advantage of cognitive biases. Some common techniques include framing things to consumers in a positive or negative light, getting users to prioritize immediate consequences over future consequences, overwhelming users with choices and decisions, selecting harmful defaults, and making users over-reliant on provided information [9, 10].

D. Contextual Integrity

The framework of Contextual Integrity [18] posits that people have a right to privacy and a right to live in a world where our expectations about the flow of personal information are met. The framework argues that privacy can be ascertained by using flows of information and characterizing the appropriateness of these flows. Appropriate information flows conform with expectations and norms based on the its contextual information. These flows can be modeled using the following parameters: data subject, sender, recipient, information type, and transmission principle. People are typically upset about their privacy being violated if their expectations are subverted or a violation of these norms occurs. One of the main arguments is that online social and informational norms should not be separate from real life. There is a need to identify appropriate contexts and norms online. For example, consumers and commercial information should be protected in the case of online transactions.

Contexts are not formally defined, rather they are abstract representations of social structures found in daily life. One of the reasons behind this is that conceptions of privacy are based on ethical concerns that evolve over time. These contexts represent how we interact with each other based on capacities structured by social spheres. In terms of their key characteristics, there can be a great amount of variability. Additionally, overlapping contexts could involve conflicts in the expectations of data flows. In online settings, privacy policies aim to inform users of the various methods and reasons for which data is collected and processed.

E. Privacy Regulation Frameworks

Using the structures and arguments provided by the Contextual Integrity framework of privacy, several regulatory organizations have created frameworks to protect user privacy. The General Data Protection Regulation (GDPR) is an EU law which provides rules for how organizations and companies handle data privacy [32]. The central idea behind its existence relies on the concept that everyone has the right to their own private affairs with others respecting this boundary. The GDPR provides each person with certain rights of their own data.

These include: the right to be informed, the right of access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object, and rights in relation to automated decision making and profiling. Below are some of the main tenets of the GDPR.

The California Consumer Privacy Act (CCPA) secures privacy rights for California residents, the main tenets of which are 1) the right to know about the personal information a business collects about them and how it is used and shared, 2) the right to delete their personal information, 3) the right to opt-out of the sale of their personal data, and 4) the right to non-discrimination when exercising their CCPA rights.

Overall, these data privacy regulations aim to increase the transparency of the data practices employed by companies that process data. However, these regulations still fall short. Privacy policies and data practice disclosures are growing in both length and complexity in an attempt to cover all legal bases [11]. These privacy laws also lack enforceable standardization and transparency requirements despite the standardized mediums of data collection by companies (HTTP(S) requests, HTTP cookies, REST APIs, etc.).

III. Motivations and Arguments

A. Notice and Consent

As it stands, the current form of notice and consent is broken. No one reads privacy policies because they are long and uninformative. The introduction of third party data partners exacerbates this problem exponentially, by directing users to read the data policies of each third party. For example, while Funny Weather (an entertaining weather mobile app), does not sell your location data directly to government and military agencies, it sells your data to Predicio (a location broker). This data aggregator resells your data to Venntel, an organization with customers such as the Federal Bureau of Investigation (FBI) and the U.S. Immigration and Customs Enforcement (ICE) [33]. Predicio alone shares data with 519 partners, many of which being data brokers (e.g., Comscore,) which share data with another set of partners. McDonald et al. [22] estimate that the average U.S. internet user must spend 76 work days reading privacy policies each year. However, the original assumption that a user encounters 1462 unique privacy policies per year was computed using the number of unique websites visited each year. If a single website or mobile app uses Predicio, they will have to read an additional 519 privacy policies and the privacy policies used by each of Predicio's partners' partners, taking at least 27 work days. Third-party data transmissions need to be made transparent, or users will never understand the implications of consenting.

The system of notice and consent is outdated and unscalable to today's' numerous amounts of software and data processing systems. Hiring lawyers to draft legal documentation for privacy practices which can be more effectively disclosed to users is an unnecessary additional expense. As the approach of notice and consent is still widely used, recent privacy regulations have been ineffective in fixing the broadness and self-contradictory nature of statements in privacy policies.

B. Natural Language Processing

The current state of the art approach for automatically extracting and regulating claims in privacy policies leverages natural language processing techniques and language models. Several automated privacy tools have been developed to interpret privacy policies, automate consent mechanisms, and monitor policy compliance, the foundations of which have been rooted in the contextual integrity principle presented by Nissenbaum *et al.* [18].

Many solutions leveraging machine learning have been proposed for improving the interpretability and presentation of privacy policies. PI-Extract [2] automatically extracts privacy practices, highlighting and describing the data practices. Polisis [15] uses neural network language models trained on many privacy policies to allow users to query the application about various privacy practices. This application is presented both in the form of a visualization tool and a chatbot. Opt-Out Easy [34] extracts and classifies opt-out choices in the form of a browser extension. PrivacyCheck [35] extracts and summarizes information from privacy policies and presents information about data practices to its users by answering a list of 20 privacy questions.

PurPliance [36] automates the regulation and compliance extraction from mobile application privacy policies. Other works include PolicyLint [37] and PoliCheck [38], which similarly construct data flows from privacy policies and cross-check these data flows with actual observed data practices. Another work trains and evaluates a named entity recognition (NER) model specifically for third-party entities [39]. These systems all rely on sentence-level NLP, NER, and dependency trees to represent the collection and sharing of data.

Cookies and fingerprinting are commonly used methods for tracking individuals' activities online. WhoTracks.Me [40] performs measurements of online tracking behavior by using 5 million users who enabled the Ghostery [41] ad-blocking browser extension. They aggregate and present the data for each category of cookie across the web. CookieBlock [42] automatically classifies and removes cookies using a random forest model trained on cookie metadata.

However, these ML-based approaches encounter a number of challenges. For example, automatically extracting data flows using natural language models can encode inaccuracies and errors. The best language models achieve only up to around 80% accuracy. Additionally, combining language models with button detection and page segmentation systems can further reduce the end-to-end performance of automated privacy tools. As privacy policies often have vague, inconsistent, or self-contradictory statements, annotating large datasets of privacy statements is an expensive and time-consuming task (Fig. 1).

C. Mandated Data Disclosures

Privacy is something that can be standardized/automated, especially if companies already have privacy engineers. However, there are conflicts of interest and little incentive for companies to take the initiative on standardizing the disclosure of privacy practices. Companies such as Onetrust

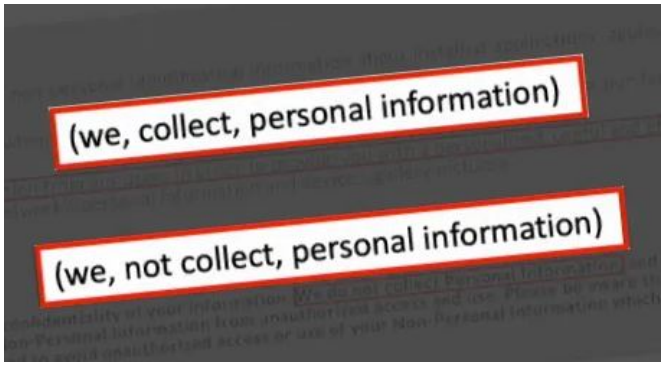

Fig. 1. Contradictory statements found within the same privacy policy [37]

(see Fig. 2) and Cookiebot have attempted to standardize cookie privacy, but these solutions still employ dark patterns which nudge users into making decisions harmful to their privacy. In order to create privacy standards, privacy regulators can take inspiration from other regulatory administrations such as the Internal Revenue Service (IRS) or the Food and Drug Administration (FDA). Income taxes and financial disclosures can be standardized with forms financial forms, and in principle, data privacy could also be standardized with more fine-grained methods of disclosure. This can be done by creating a large table of data types, collection purposes, data usages, medium of collection, etc. An sample data transaction disclosure can be seen in **??**. If fine-grained disclosures of each data transaction made on the user's data were mandated, this would allow for privacy researchers and companies to develop tools that can easily parse this standardized data disclosure format. Regulation, summarization, and visualization could all be automated by leveraging this standardized disclosure method. The main added requirements for data processors would include the following: 1) data processors must disclose the details of each data flow involving a user's information, 2) data processors must disclose the name of each client or partner which retrieves or sends user information, and 3) data processors must conform to the data disclosure standards and provide an accurate representation of their actual data practices.

IV. A Model of Internet Privacy

Realistically, constructing context for privacy requires a formalization with more than just the five parameters derived in the model of contextual integrity by Nissenbaum et al. [18]. It requires the ability for users, regulators, or privacy tools to parse fine-grained details about data types, their usage purposes, the collection or sharing medium, the transmission details, and more. To address this issue, we create a potentially exhaustive set of parameters for regulators to consider as a standard and describe how to process this information.

For example, building a context requires knowledge of the data owners, data senders, and data recipients. They can be companies, organizations, data brokers, end users, and more. These agents can eventually be associated with a particular level of trustworthiness, roles, privacy inclinations, and motivations behind their actions. This information would have to be generated by holistic analyses of data flows and companies.

There are also many specific pieces of data. These can include personal information, location data, photos, high dimensional representational embeddings, inferred interests, clicks, and more. Data has different types and can be used for many different purposes. Data is often collected with the intention of re-transmission to one or more third party entities. As a result, the context around a piece of data can grow exponentially, becoming complicated very quickly. Once data is shared with a third party, this data is now subject to any additional agreements or re-transmission principles found in the third party's privacy policy.

Context around data privacy needs to specify the usage/purpose behind a particular data transmission. Users may be upset if their location data is used for surveillance purposes or made available to government entities, however, they may be ambivalent if their location data is used for targeted advertising purposes.

Each data flow involves a data owner, data sender, and a data recipient. It also includes the data being transmitted. Each data flow also has a particular medium through which it is exchanged. Additionally, data transmissions are associated with an activity type. Data transmissions can also be continuous, one-time, or event-based. Finally, the value of each individual piece of data should be constructed using the available context surrounding each data flow. Data transmissions can be used to model disconnects between user expectations and privacy

Cookie Name	Domains	Lifespan	Hostname	Category	Description
_ga	preferencechoice.com	PERSISTENT	preferencechoice.com	Analytics	This cookie name is associated with Google Universal Analytic...
lang	cisco.com, onetrust.com	SESSION	linkedin.com	Targeting	This domain is owned by LinkedIn, the business networking pl...
_gclxxxx	vendorpedia.com	PERSISTENT	vendorpedia.com	Analytics	Google conversion tracking cookie
pctrk	onetrust.com	PERSISTENT	my.onetrust.com	Strictly Necessary	Used to count page views by unauthenticated users against lic...
OptanonConsent	preferencechoice.com	PERSISTENT	preferencechoice.com	Strictly Necessary	This cookie is set by the cookie compliance solution from One...

Fig. 2. An example cookie table created for a company and provided to users by OneTrust.

policies. They can also be used to model dissonance between privacy policies and actual implementations.

Failures in the notice and consent model occur whenever the user's expectation of transmissions fail to align with the actual transmissions. Privacy violations occur whenever the transmissions detailed in privacy policies and notices fail to align with actual transmissions. Both consent failures and privacy violations can occur whenever a transmission is omitted or a transmission's properties differ.

A. Usability

While this model would function well for automated regulation in a given privacy policy and service, users will be unable to reap the benefits of this formal model. It is critical that users are made aware of the entire network of data vendors/customers and notified of the most concerning entities handling their data. As such, the information captured by the model needs to be distilled into easily understandable formats for users while maintaining the accuracy of the formalization.

For example, privacy regulation can take inspiration from regulations standardized in industries such as the food & drug industry. Similar to the concept of nutrition labels or side effect disclosures for food and drugs, researchers have also developed an IoT privacy label [17] to better inform users thinking of buying hardware products which collect data. Mobile app and browser extension permissions have implemented this approach successfully, distilling information and significantly improving usability and presentation. Just as people make health conscious decisions based on ingredients and nutrition information, if people are made aware of all the data collection and tracking mechanisms, they would more carefully consider their privacy.

Fig. 3. A potential design for IoT privacy labels [17]

V. Discussion

A. Dilemmas

One common argument made against targeted advertising and data privacy advocation is the possibility that prioritizing privacy would hurt the economy and slow growth. With companies adopting the surveillance capitalism business model, this business model of aggregating and sharing data with few constraints has allowed for the rapid creation of many machine learning models. The targeted advertising industry creates many beneficial services which allow marketers to efficiently and effectively advertise products to their core demographics/profiles. Banning personalized advertising would certainly negatively impact the economy [43] by causing customer acquisition and advertising costs to increase in cost. This would increase expenses for both companies and nonprofit organizations. Though a dilemma emerges through this – targeted advertising in its current form encourages the promiscuous collection and sharing of vast quantities of data. Furthermore, the same data collected for targeted advertising purposes could be used by law enforcement or government entities to monitor personal activity. The mishandling and misuse of data is an inevitable byproduct stemming from the widespread proliferation of targeted advertising. On the more sensitive spectrum of data types, location data, face data, social networks, browsing history, and transaction history could be used to automate surveillance and stalking activities. Thus privacy regulators have the difficult task of striking a compromise in the tradeoff between customer acquisition and privacy. Privacy regulators should still uphold people's right to privacy by mandating companies to be transparent and manipulative with their data privacy practices and disclosures.

TABLE I
An example parameter list for a standardized disclosure of data flows. Each data flow is represented by a column.

Owner	Sender	Recipient	Data Name	Data Type	Data Usages	Data Expiration	Medium	Activity	Continuous	Purpose
Brian Tang	Brian Tang	Google	GPS coordinates	Location	Display location, recommend places	Never	Mobile app	Map	Continuous	Functionality
Brian Tang	Brian Tang	Facebook	Face picture	Image	Tag photo	Never	Website upload	Profile update	1-time	Functionality
Brian Tang	Brian Tang	Facebook	Post click	UI Interaction	Attribute sponsored links	30 days	Website cookie	Feed browsing	1-time	Advertising
Brian Tang	Facebook	Microsoft	Post click	UI Interaction	Never	Attribute sponsored links	REST API	Data sale	1-time	Data aggregation

B. Limitations

Our calls for increased standardization and regulation in online data privacy may be met with resistance. Particularly, our approach would put an increased burden on developers and engineers working for data processors. However, the process of disclosing the information related to data practices can be automated, especially since data, software, websites, and companies all already have standard protocols for moving data around (e.g., cookies, HTTP/S requests, csvs, REST APIs). We argue that these information types can be automatically scraped or monitored and inserted into a table or a form to be retrieved from the company website. The current systems places an unreasonable burden of understanding privacy policies onto users. Why should users bear the cost of reading numerous privacy policies when companies have the manpower to improve the transparency and honesty of their privacy disclosures?

VI. Conclusion

As data privacy lingers as a main concern on the minds of users and regulators, researchers are working to create privacy tools to improve users' experiences. We discuss how the current "notice and consent" approach to data privacy is unfair to users and unscalable in the modern culture of sharing data. Rather than providing users with long privacy policies, the disclosure of privacy practices should be standardized, and each data transmission should be recorded as an entry in a table. This table can be easily parsed as CSV or JSON files, and privacy tools/regulators no longer have to rely on the error-prone approach of using natural language models to interpret vague privacy policies. Through this standardized framework, researchers can focus more on the usability aspect by creating intuitive summarizations and visualizations of data practices rather than demystifying opaque data practices. This approach would save time and money for both businesses and privacy-conscious consumers.

References

[1] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[2] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin, "Automated extraction and presentation of data practices in privacy policies.," *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 2, pp. 88–110, 2021.

[3] A. Mathur *et al.*, "Dark patterns at scale: Findings from a crawl of 11k shopping websites," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–32, 2019.

[4] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher, "Tales from the dark side: Privacy dark strategies and privacy dark patterns.," *Proc. Priv. Enhancing Technol.*, vol. 2016, no. 4, pp. 237–254, 2016.

[5] *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information*, [Online; accessed 31. Jan. 2022], Aug. 2020. [Online]. Available: https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information.

[6] *64% of Americans Don't Know What to Do After a Data Breach — Do You? (Survey)*, [Online; accessed 31. Jan. 2022], Jan. 2022. [Online]. Available: https://www.varonis.com/blog/data-breach-literacy-survey.

[7] E. Goitein, "The government can't seize your digital data. Except by buying it.," *Washington Post*, Apr. 2021, ISSN: 0190-8286. [Online]. Available: https://www.washingtonpost.com/outlook/2021/04/26/constitution-digital-privacy-loopholes-purchases.

[8] *Beware Of Privacy-Policy Loopholes! - Cybernetic GI*, [Online; accessed 31. Jan. 2022], May 2020. [Online]. Available: https://www.cyberneticgi.com/2020/04/30/beware-of-privacy-policy-loopholes.

[9] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, 2015. DOI: 10.1126/science.aaa1465. eprint: https://www.science.org/doi/pdf/10.1126/science.aaa1465. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaa1465.

[10] A. E. Waldman, "Cognitive biases, dark patterns, and the 'privacy paradox'," *Current opinion in psychology*, vol. 31, pp. 105–109, 2020.

[11] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the gdpr," *arXiv preprint arXiv:1809.08396*, 2018.

[12] E. Roth, "Google abandons FLoC, introduces Topics API to replace tracking cookies," *Verge*, Jan. 2022. [Online]. Available: https://www.theverge.com/2022/1/25/22900567/google-floc-abandon-topics-api-cookies-tracking.

[13] google, *ads-privacy*, [Online; accessed 31. Jan. 2022], Jan. 2022. [Online]. Available: https://github.com/google/ads-privacy/blob/master/proposals/FLoC/FLOC-Whitepaper-Google.pdf.

[14] *DigitalAdvertisingAlliance.org*, [Online; accessed 31. Jan. 2022], Jan. 2022. [Online]. Available: https://digitaladvertisingalliance.org.

[15] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th {USENIX} security symposium ({USENIX} security 18)*, 2018, pp. 531–548.

[16] R. Khandelwal, T. Linden, H. Harkous, and K. Fawaz, "Prisec: A privacy settings enforcement controller," in *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.

[17] P. Emami-Naeini, Y. Agarwal, L. F. Cranor, and H. Hibshi, "Ask the experts: What should be on an iot privacy and security label?" In *2020 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2020, pp. 447–464.

[18] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[19] M. Steinberger, "Does palantir see too much," *New York Times Magazine*, 2020.

[20] R. H. Sloan and R. Warner, "Beyond notice and choice: Privacy, norms, and consent," *J. High Tech. L.*, vol. 14, p. 370, 2014.

[21] U. Deloitte, "Global mobile consumer survey: Us edition," *Deloitte US*, 2016.

[22] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Isjlp*, vol. 4, p. 543, 2008.

[23] S. Zuboff, "Big other: Surveillance capitalism and the prospects of an information civilization," *Journal of information technology*, vol. 30, no. 1, pp. 75–89, 2015.

[24] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.

[25] K. Litman-Navarro, "Opinion | We Read 150 Privacy Policies. They Were an Incomprehensible Disaster.," *N.Y. Times*, Jun. 2019, ISSN: 0362-4331. [Online]. Available: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html.

[26] *It's Not You; Privacy Policies Are Difficult to Read | Common Sense Education*, [Online; accessed 23. Feb. 2022], Sep. 2020. [Online]. Available: https://www.commonsense.org/education/articles/its-not-you-privacy-policies-are-difficult-to-read.

[27] D. Spagnuelo, C. Bartolini, and G. Lenzini, "Metrics for transparency," in *Data Privacy Management and Security Assurance*, Springer, 2016, pp. 3–18.

[28] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon," *Computers & security*, vol. 64, pp. 122–134, 2017.

[29] D. J. Solove, "The myth of the privacy paradox," *Geo. Wash. L. Rev.*, vol. 89, p. 1, 2021.

[30] T. Morey, T. Forbath, and A. Schoop, "Customer data: Designing for transparency and trust," *Harvard Business Review*, vol. 93, no. 5, pp. 96–105, 2015.

[31] A. Acquisti *et al.*, "Nudges for privacy and security: Understanding and assisting users' choices online," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–41, 2017.

[32] "2018 reform of eu data protection rules," European Commission. (May 25, 2018), [Online]. Available: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf (visited on 06/17/2019).

[33] *My Phone Was Spying on Me, so I Tracked Down the Surveillants*, [Online; accessed 19. Feb. 2022], Feb. 2022. [Online]. Available: https://nrkbeta.no/2020/12/03/my-phone-was-spying-on-me-so-i-tracked-down-the-surveillants.

[34] V. Bannihatti Kumar *et al.*, "Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text," in *Proceedings of The Web Conference 2020*, 2020, pp. 1943–1954.

[35] R. N. Zaeem, R. L. German, and K. S. Barber, "Privacycheck: Automatic summarization of privacy policies using data mining," *ACM Transactions on Internet Technology (TOIT)*, vol. 18, no. 4, pp. 1–18, 2018.

[36] D. Bui, Y. Yao, K. G. Shin, J.-M. Choi, and J. Shin, "Consistency analysis of data-usage purposes in mobile apps," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2824–2843.

[37] B. Andow *et al.*, "{Policylint}: Investigating internal privacy policy contradictions on google play," in *28th USENIX security symposium (USENIX security 19)*, 2019, pp. 585–602.

[38] B. Andow *et al.*, "Actions speak louder than words:{entity-sensitive} privacy policy and data flow analysis with {policheck}," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 985–1002.

[39] M. B. Hosseini, K. Pragyan, I. Reyes, and S. Egelman, "Identifying and classifying third-party entities in natural language privacy policies," in *Proceedings of the Second Workshop on Privacy in NLP*, 2020, pp. 18–27.

[40] A. Karaj, S. Macbeth, R. Berson, and J. M. Pujol, "Whotracks. me: Shedding light on the opaque world of online tracking," *arXiv preprint arXiv:1804.08959*, 2018.

[41] *Ghostery*, [Online; accessed 19. Feb. 2022], Feb. 2022. [Online]. Available: https://www.ghostery.com.

[42] D. Bollinger, K. Kubicek, C. Cotrini, and D. Basin, "Automating cookie consent and GDPR violation detection," in *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA: USENIX Association, Aug. 2022, TBA, ISBN: TBA. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger.

[43] D. C. Benjamin Mueller, "The Value of Personalized Advertising in Europe," *Information Technology and Innovation Foundation*, Nov. 2021. [Online]. Available: https://itif.org/publications/2021/11/22/value-personalized-advertising-europe.