The Realities of Disclosure Risks in the Age of Dark Patterns and Big Data

Ramon Abraham A. Sarmiento

Abstract:

This proposal aims to explore the realities of disclosure risks in the current big data landscape, examining implications for individuals, society, and the evolving ethical landscape. The pervasive use of dark patterns in data collection has sparked significant ethical and legal debates on balancing confidentiality & privacy obligations with the need for precise research data. Thus, this paper seeks to shed light on the legal and ethical dilemmas of these disclosure risks arising from the intersection of data privacy, statistical data usage, and the employment of dark patterns, such as the possible insufficiency of existing data protection measures and their possible obsolesce in the age of Big Data. It also presents a use case that assesses the disclosure risk of Philippine datasets using two key measures: the Risk of Reidentification (RRI) and k-anonymity.

1. Introduction:

In 2012, Steve Lohr of the New York Times welcomed the general public to the age of big data (Lohr 2012).[1] He touted many of the advances and changes that we barely notice today, such as the rapid gathering of information by social networks and their processing of said information in new ways to gain increasingly valuable insights. Lohr did not mention that some of this gathering of data may be harmful such as emergence of dark patterns for the massive collection of digital information or the empowerment of AI powered surveillance states (Beraja et. al. 2023). The annual Domo "Data Never Sleeps Infographic" illustrates this growth in data generation and gathering, showing internet users have increased to 5.2 billion in 2023 from 2.1 billion in 2011.[2] The emergence of big data analytics has increased the effort put into collecting and processing vast amounts of data (Bahga & Madisetti, 2016).

Even prior to the emergence of big data into the public eye, Solove (2004) already pointed out that we leave traces of information whatever we do and wherever we go. The emergence of big data, means that it is more likely that these information traces will be collected, shared, processed and in a sense monetized, which makes preserving privacy much harder. Thus, it is no wonder that researchers have pointed out the high likelihood that big data will take a role in changing social and economic policy and research (Mazzie & Noble 2019).

The explosion of data generation and collection has changed the validity of the current frameworks that measure and minimize disclosure risks. The 2016 Federal Big Data and Research and Development Strategic Plan (2016 Federal Big Data Plan) of the U.S. Networking and Information Technology Research and Development Program (NITRDP) explicitly stated that: *"Emerging Big Data*

[1] Francis X. Diebold made first academic reference to Big Data in a title or abstract in the statistics or econometrics, literatures in his paper "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting," presented at the Eighth World Congress of the Econometric Society in Seattle in August 2000, and subsequently published in 2003. Diebold traces the origins of the term "Big Data" in his paper "On the Origin(s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline".

[2] https://web-assets.domo.com/blog/wp-content/uploads/2023/12/23-dns11-FINAL-1.png

technologies hold great promise for society, but also present new challenges to the ethical use of data, analyses, and results, and to the privacy and security of data. The solutions and approaches needed to address these challenges require deep attention and will have a major impact on the ability to access, share, and use Big Data."[3] The truth is that much of this collection and processing has been benign but there are documented cases of both data breaches[4] and misuses through dark patterns among others.[5] Notwithstanding the potential and documented privacy harms, big data still holds immense potential for societal progress with many studies on use cases across the public and private sector, such as healthcare (Cremin et. al. 2022), finance (Hasan et. al. 2020) and even counterterrorism (Reilly 2015).

There are still numerous concerns regarding data privacy and ethical use of statistical data, particularly disclosure risk and its attendant harms. Hotz et. al. (2022) broadly defined "disclosure risk" as the probability that presumed private and specified information about a particular data subject in a particular database will be obtained by an unauthorized party and associated with the data subject. Disclosure risks may also inform the willingness of individuals and business to provide sensitive information in support of research and statistical initiatives (Kern et. al. 2018, and Hotz & Slanchev 2017).[7] The use of dark patterns in digital interfaces exacerbates these concerns by manipulating user behavior without their informed consent, which in many instances makes these users share personal information against their best interests.

Big data fuels rapid datafication which in turn fuels the current business models of data brokers driving yet further datafication (Ruschemeier 2023). Moreover, the greater demand for data, puts

---

[3] Available at https://www.nitrd.gov/pubs/bigdatardstrategicplan.pdf; Hereafter referred to as the NITRDP BDSP; Last visited 28 March 2024

[4] https://privacyrights.org/resources/united-states-data-breach-notification-united-states-2023-report

[5] See Press Release of the European Commission "Consumer protection: manipulative online practices found on 148 out of 399 online shops screened" 30 January 2023, available at https://ec.europa.eu/commission/presscorner/detail/en/ip_23_418

[7] See also Principle 6 – Confidentiality - I. Objective of United Nations Fundamental Principles of Official Statistics Implementation Guidelines, which states: "*A fundamental requirement for official statistics is confidence and acceptance of public. Accurate and timely data are reliant on public goodwill and cooperation no matter if their participation is facultative or if it is based on compulsory response. In order to maintain the trust of respondents it is the utmost concern of official statistics, to secure the privacy of data providers (like households or enterprises) by assuring that no data is published that might be related to an identifiable person or business. At the same time this guarantees quality by avoiding loss of accurate data. Confidentiality protection is supposed to be implemented on each level of the statistical process from the preparation of surveys up to the dissemination of statistical products.*"
Available at https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf

increasing pressure on data custodians to make databases available to outsiders, even if there is a lack of resources, tools, or expertise needed to do so safely (*Schmutte and Vilhuber 2020)*. Nissenbaum (2009) observed that "*Privacy has been one of the most enduring social issues associated with digital electronic information technologies*." Advances in big data and the prevalence of dark patterns add new complex elements to this observation.

2. Significance

This proposal seeks to delve into these intricate ethical quandaries in the current big data landscape to provide an overview of the possible insufficiency of existing data protection measures and their possible obsolesce in the age of big data. The privacy impact of dark patterns has been studied alongside the consumer impacts such as a willingness to pay higher prices (Blake et. al. 2021). Big data analysis of digital systems that individuals interact with reveal their individual idiosyncrasies and the patterns in the groups to which they belong, revealing weaknesses and dispositions that individuals themselves cannot see (Susser et. al. 2018).

As big data analytical methods progress and the amount of data continues to increase, it becomes more difficult to preserve the required level of privacy and availability in data sets made available for research, while still providing the required level of utility for the stated research goals. The application of differential privacy or other privacy protections involves trade-offs between privacy and accuracy/utility (Acquisti et. al. 2016, Oberski & Kreuter 2020, and Groshen & Goroff 2023). One of the reasons that confidentiality and anonymity are valued is due to the notion that anonymous respondents are more forthcoming in their responses to questions asked of them on a survey instrument or other research (Stoughton et. al. 2015). Advances in probabilistic record linkages and other big data processes could render moot our existing technical privacy preservation methods, despite the consistent calls for reviews of the current legal and technological frameworks for privacy preservation. Ultimately, the appropriate value plausible deniability for a published data set is not a mathematical question, but rather a policy

decision for the publisher to make and maintain (Groshen & Goroff 2023). Thus, a closer look into the interactions between big data, dark patterns, data brokers, disclosure risk and their ethical frameworks is in order.

3. Limitations and Research Challenges

This study will take a broad overview of the big data landscape and the attendant disclosure risks due to dark patterns. Only estimations of the likelihood of this reversal of de-identification shall be provided due to ethical and legal concerns regarding the possession of de-anonymized data and commercially available data sets where the original basis for processing is not readily apparent. This paper will not seek to prove technical weaknesses in existing privacy preservation techniques as applied to actual datasets but rather look into the possible gaps in existing ethical frameworks that increase disclosure risks.

Due to the nature of data, it is nearly impossible to tell at the level of commercially available data sets, which of the component data was sourced from dark patterns due to the lack of any standard disclosure notices on the sale of these data sets and lack of comprehensive rules regulating data brokers. Nonetheless, there are ample studies documenting the effectiveness of dark patterns in collecting personal information such that a specific terms such as Privacy Zuckering have been coined.[8] This also begs the question on how to treat and possibly regulate data sets that are composed both of data sourced from dark patterns, and data that was sourced legitimately e.g. scrapped from public databases. Many surveys show people have a strong distaste for the sale of their data and given the openness of the sale of these data sets, it would be fair to conclude that a sizable portion was sourced using dark patterns. Absent more detailed studies on the prevalence of data sets assembled by privacy infringing dark patterns, commercially data sets that can be purchased with little or no limitations, are a reasonable proxy for illustrating disclosure risks in relation to the prevalence of dark patterns.

[8] Tim Jones, “Facebook’s "evil interfaces", available at https://www.eff.org/deeplinks/2010/04/facebooks-evil-interfaces; last visited 03 April 2024
See also Mohit, Privacy Zuckering: Deceiving your privacy by Design, 11 April 2017, available at https://medium.com/@mohityadav0493/privacy-zuckering-deceiving-your-privacy-by-design-d41b6263b564; last visited 03 April 2024

Lastly, determining the impacts of recent regulations on dark patterns presents similar challenges to those identified for research on the economic impact of the GDPR by Johnson (2023) such as: (1) finding a suitable control group, (2) variations in enforcement, and compliance, due to the staggered nature of regulations on dark patterns, and (3) impact of dark pattern regulations on data observability. There is an added difficulty in that many studies of dark patterns do not make a separate assessment of the privacy impacts from other potential consumer harms.

4. Literature Review

4.1 *What is a Dark Pattern?*

Harry Brignull coined the term dark patterns in 2010.[9] He defined dark patterns as "*A Dark Pattern is a manipulative or deceptive trick in software that gets users to complete an action that they would not otherwise have done, if they had understood it or had a choice at the time. For example if you have a button that functions as a "Yes" when clicked, but through the use of placement, colour and trick wording, it appears to say "No", then many users are going to be caught out."*[10] Mathur et. al. (2021) in their review of literature on dark patterns have identified at least 19 different definitions of dark patterns. The literature on dark patterns shares similarities and often intersects with the studies made by researchers on online manipulation (Susser et al., 2019), malicious interfaces, nudges, surveillance capitalism[11] and UX design (Kowalcyzk et. al. 2023). There are numerous definitions and conceptual frameworks of dark patterns (Mathur 2021, Kollmer & Eckhardt 2022, and Gray et. al. 2023). Beyond, these definitions and conceptual frameworks, the main purpose of a dark pattern is to manipulate people to pay for something they otherwise would not purchase or surrender personal information they would

[9] See https://www.deceptive.design/; see also Harry Brignull, Bringing Dark Patterns to Light; available at https://harrybr.medium.com/bringing-dark-patterns-to-light-d86f24224ebf (last visited on 24 March 2024).

[10] Brignull, Harry, Bringing Dark Patterns to Light, 07 June 2021; Definition is from a speech Harry Brignull gave during Federal Trade Commission's Dark Patterns workshop on April 29, 2021, available at https://harrybr.medium.com/bringing-dark-patterns-to-light-d86f24224ebf;

[11] See https://guides.libraries.psu.edu/berks/darkpatterns; last visited on 06 April 2024

otherwise keep confidential (Luguri & Strahilevitz, 2021). For this paper, we focus on the surrender of personal information.

Kelly & Burkell (2023) have found that privacy dark patterns prevent users from making conscious, informed decisions about the management of their personal data which in turn exposes them to a number of risks and harms, including cyberstalking, identity theft and reputational damage. A large-scale experiment conducted by on census-weighted samples of American adults showed that mild dark patterns more than doubled the percentage of consumers who signed up for a dubious identity theft protection service, and aggressive dark patterns nearly quadrupled the percentage of consumers signing up (Luguri & Strahilevitz 2021). The dynamic, interactive, intrusive, and adaptive choice architectures of digital platforms are the ideal medium for leveraging these insights into our decision-making vulnerabilities (Susser et al., 2019). Moreover, simply through frequent use and habituation, digital technologies become invisible and thus ideal vehicles for manipulation (Susser et al., 2019). Social networking sites were also found to have deployed multiple dark patterns that complemented and reinforced one another. (Kelly & Burkell 2023).

Dark patterns pervasiveness and impacts may be still understated. There has been little attention paid to their presence on the internet of things (IOT) as most of the studies on dark patterns have focused on websites and applications. Owens et. al. (2022) have found the presence of dark patterns in voice interfaces. Kowalczyk et. al. (2023) found at least 3 unique dark patterns in all 57 Internet-of-Things (IoT) devices studied and on each IOT devices contained an average of over 25 dark patterns. Kowalcyzk et. al. (2023) surmised that many IoT devices show dark patterns multiple times and in large numbers. Luguri & Strahilevitz (2021) suspect that the effectiveness of dark patterns has been replicated by social scientists in technology and e-commerce companies, because the internal, proprietary research suggests dark patterns generate profits for the firms that employ them. This effectivity and secrecy seems to be reinforced by the Facebook documents leaked by former Facebook data scientist Frances Haugen or the

aptly named "Facebook Papers".[12] Kallioniemi (2022) reports that these leaks caused many corporations to take measures to prevent events like this from happening again and have caused the companies to close up and spy on their own employees even more.

4.2 Dark Patterns and Privacy

Rubenstein (2013) argued that the advancing wave of big data will overwhelm the informed choice model in the European Union Data Protection Directive 95/46 EC (DPD).[13] Liu et. al. (2023) have pointed out that consumer privacy faces an unprecedented challenge by this massive collection of consumer data by digital platforms, noting that a considerable portion of these digital platforms have adopted dark patterns. Bösch et al. (2016) presented a categorization of privacy-specific dark patterns as 'Dark Strategies' as analogous to existing privacy design strategies conceptualized by Jank-Henk Hoepman.[14] Bösch et al. (2016) then enumerated several privacy dark patterns such as Privacy Zuckering, Bad Defaults, Forced Registration, Hidden Legalese Stipulations, Immortal Accounts, Address Book Leeching and Shadow User Profiles. By their very nature, these dark patterns circumvent the ability of an individual to make an informed choice as to their personal information.

Numerous studies have shown the effectiveness of dark patterns in manipulating individuals into surrendering their personal information. Often using content analysis, researchers evaluate interface elements such as cookies and shown whether they contained dark patterns or not (Gray et. al. 2023 Mapping). Famously, a Norwegian study concluded that dark patterns and privacy intrusive defaults, would nudge users of Facebook and Google, and to a lesser degree Windows 10, toward the least privacy friendly options to an unethical degree (Forbrukerrådet, 2018a). This Norwegian study also found that Google and Facebook only gave their users an illusion of control of their personal data, which susceptible

[12] See https://facebookpapers.com/; last visited 06 April 2024

[13] Directive (EC) 95/46 of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data
[1995] OJ L281/31; available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046;
This Directive is the predecessor of the GDPR, the latter also has the informed choice model.

[14] See Jank-Henk Hoepman, Privacy Design Strategies (The Little Blue Book), version as of 19 April 2022 available at https://www.cs.ru.nl/~jhh/publications/pds-booklet.pdf; last visited 03 April 2024

to take more risks when disclosing sensitive personal information.[15] Labelling it as the control paradox (Forbrukerrådet, 2018a).[16] In a separate 2018 study, the Forbrukerrådet (2018b) claimed that Google deceived consumers into being tracked when they use Google services, through a variety of techniques including dark patterns.

Utz et. al. (2019) in their analysis of a random sample of 1,000 Consent Management Platforms (CMPs) found that at least 57.4% used dark patterns to nudge users to select privacy-unfriendly options, and that 95.8% provide either no consent choice or confirmation only. This is like the study conducted by Nouwens et. al, (2020) where they found less than 12% of the design and text of the CMPs used by the top 10,000 websites in the UK to be compliant with EU law. Another study by Matte et. al. (2019) detected at least one suspected violation in the way consent is determined, asked, or complied for 54% of websites using AB Europe's Transparency and Consent Framework (TCF) cookie banners.. They further conclude that consent notices with a meaningful adherence to the GDPR would lead to less than 0.1 % of users actively consenting to the use of third-party cookies. Soe et. al. (2020) found dark patterns in 297 out of 300 data collection consent notices from news outlet websites that are built to ensure compliance with GDPR, this included 175 notices not indicating the purpose of the cookies used. Mathur et. al. (2019) found that 1,818 dark patterns were in 11.1% or 1,254 of the 11,000 shopping websites, and the more popular shopping website were more likely to have dark patterns. Utz et. al (2019) conclude that many current cookie notice implementations under the GDPR do not offer meaningful choice to consumers.

Tahaei et. al (2022) showed that 77% of free apps have an ad library, which in turn allows advertisers to collect and track user data across a wide variety of applications. The defaults configurations of these ad networks are set to maximize data collection and may have dark patterns (Mhaidli et. al. 2019).

[15] The study found that Facebook users, even after going through the extra effort of changing their settings with the intention of using their privacy rights, are not given a substantial choice with respect to how their personal data is collected and used. With respect to Google users, the privacy dashboard discourages users from changing or taking control of the settings or delete bulks of data, due to the overwhelming amount of granular choices to micromanage.

[16] Citing "Gone in 15 Seconds: The Limits of Privacy Transparency and Control" https://www.computer.org/csdl/mags/sp/2013/04/msp2013040072.html

In fact, a review of 4 popular ad networks showed the use of dark patterns to nudge developers to choose personalized ads and share more data, rather than giving developers a chance to make informed choices (Tahaei and Vaniea 2021).

Di Geronimo et. al. (2020) found that 95% of 240 popular mobile applications contained one or more dark patterns and, on average. Popular applications contained an average 7.4 dark patterns. Using the qualitative expert classification study by Di Geronimo et al. (2020) for the Japanese app context, Hidaka et. al. (2023) found that 93% of 200 popular Japanese mobile applications contained one or more dark patterns, and on average they contained 3.9 dark patterns.

In a more recent study, Utz et. al. (2023) found that for websites such as Youtube, dark patterns often hide privacy-friendlier configurations. These privacy harms may not be readily apparent, since even data disclosed to these online platforms that is restricted from public view, increases the ability of sites to infer additional, sensitive information about users and to detect person-specific vulnerabilities (Susser et. al., 2019 and Kelly & Burkell 2023).

The study by Mhaidli et. al. (2019) also shows that many application developers are aware of this collection sans consent as shown in a survey of 49 United Kingdom based application developers, the results showed that 41% of respondents that ad networks collect user data without users' permission as probably/definitely true, with only 20% of respondents seeing it as false.

Leiser (2020) gives an overview of the harms and risk of dark patterns*: "The risks arising from this type of system architecture design are amplified when its sources and intentions remain hidden: when users are subjected to customized architecture and dark pattern tactics, they cease to be subject to not only regulatory scrutiny, but make decisions that go against their interest, and, on occasion, harm the entire digital ecosystem."* This includes nudging which manipulate users into disregarding their privacy and to supply more personal data than necessary (Leiser 2020). Putting an end to this collection of

personal data by digital platforms is easier said than done, due to the free services provided by these digital platforms such as email, messaging, social networking (Liu et. al. 2023).

Given the pervasiveness of these dark patterns in websites, applications and even IoT devices, it is no wonder that we cannot keep a reasonable track of where our data goes. Individuals cannot take steps to correct their data, nor rectify errors in consent which may have been caused by dark patterns. This is scale is best illustrated by a recent study conducted by Marti et. al. (2024), which found 186,892 companies sent data about 709 volunteers to Facebook. Each volunteer had an average of 2,230 companies sharing their data with Facebook and some volunteers had over 7,000 companies providing their data (Marti et. al. 2024).[19] This tracking of individuals was recently highlighted by the Court of Justice of the EU in C-252/21 Meta Platforms Ireland and Others v. Bundeskartellamt.[20] Thus, the relationship between dark patterns and disclosure risks needs to be closely examined given the prevalence of large private databases containing personal information coupled with growing sophistication of algorithms to match those databases to data released by the government and research institutions (Hotz et. al. 2022).

4.3 Disclosure Risks

Disclosure can refer to identification which is the association of a known individual with a particular microdata record or attribution which is the association of information in a micro-dataset with a particular individual. Although disclosure may seem difficult for an isolated dataset with supposedly "anonymized" data,[22] re-identification can be performed by linking public records such as voter registration records, with de-identified data (Dwork et. al. 2017). In a highly cited study that was given significant media attention, Rocher et. al. (2019) found that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. These demographic attributes are contained in

[19]The study does not "make any claims about how representative this sample is of the U.S. population as a whole, because the data came from a self-selected group of users, and because the results were not demographically adjusted. Moreover, individuals who care about privacy are surely overrepresented in this data as many were recruited through privacy concerned groups.

[20] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62021CJ0252

[22] Under legal definitions such as the GDPR, if a data set permits re-identification then it is deemed pseudonymized or de-identified (UK term, see Section 171 of the UK Data Protection Act of 2018). Recital 26 explains anonymous information. Article 4(5) of the GDPR in relation to Recital 28 explain pseudonymized information.

thousands, possibly tens of thousands of databases, many of which can be easily purchased or leased. Dark patterns heighten these disclosure risks since many people are unaware of the richness of their records as well the breadth of circulation of these records.

There are valid privacy concerns that sophisticated internet tools enable prying eyes to find even small items of information about someone (Jouhki et. al. 2016, and Nycyk 2021). Previous disclosure avoid measures relied on protecting data with noise by imagining what 'sensitive' values an attacker would want to target, attack methods and databases that would reasonably be used (Oberski & Kreuter 2020). Groshen & Goroff (2023) have pointed out that these previously adequate disclosure avoidance procedures now leave people at risk for re-identification due to the ease of finding personal information due to the internet. Researchers have demonstrated that there is still a substantial risk of re-identification even in country-scale location datasets (Farzanehfar et al. 2021). Such is the breadth of reidentification, that Henriksen-Bulmer & Sheridan (2016) did a systematic review of literature on re-identification attacks. However, not all linkage and re-identification is undesirable as linkage of person-based administrative data to survey data is used to reduce survey costs and ease respondent and interviewer burden (Sala et. al. 2014).

Dwork et al. (2017) provided an overview of re-identification and reconstruction attacks. There are numerous studies that have re-identified in whole or in part, supposedly anonymous data sets. The New York Times revealed the identity of AOL user no. 4417749, after AOL released search records of 650,000 users covering a 3 month period were released to the public.[23] Narayanan & Shmatikov (2008) used a de-anonymization methodology for sparse micro-data to successfully identify the known Netflix users in a dataset publicly released by Netflix. They uncovered apparent political preferences and other potentially sensitive information, by using the Internet Movie Database as the source of background

[23] MICHAEL BARBARO and TOM ZELLER Jr. "A Face Is Exposed for AOL Searcher No. 4417749" 09 August 2006; New York Times, available at http://www2.hawaii.edu/~strev/ICS614/materials/NYT%20-%20confidentiality%20-%20A%20Face%20is%20Exposed%20for%20AOL%20Searcher%20%202006-08-24.pdf; last visited 07 April 2024. This AOL dataset is still readily available online for free and be found after a brief search.

knowledge. Netflix was sued for this disclosure in class action suit, as the plaintiffs claimed it constituted a violated among other things, the Video Privacy Protection Act of 1988.[24] The lead plaintiff in this case, Jane Doe, claimed that Netflix's disclosure of her movie rental history and ratings has and/or will "*identify or permit inference of her sexual orientation… [which…] would negatively affect her ability to pursue her livelihood and support her family, and would hinder her and her children' ability to live peaceful lives within Plaintiff Doe's community."*[25] Henriksen-Bulmer & Sheridan (2016) note that the Netflix case was the first to prove that the scope of re-identification was possible on any released dataset and not just public datasets.

Researchers linked the Harvard Class of 2009 to the supposedly anonymous profile data for the study "Tastes, Ties and Time" (Zimmer 2010). Sweeney (2011) demonstrated the reidentification of patients in de-identified pharmaceutical marketing data using publicly available hospital discharge and ambulatory claims data, then linking it to voting list data. Culnane et. al. (2017) successfully re-identified patients in an Australian de-identified open health dataset, finding that a few mundane facts could isolate an individual and some were identified by name based on publicly available information.[26] Culnane et. al. (2019) also re-identified themselves and several others from a large passenger centric transport dataset shared by the Victorian Government for the 2018 Melbourne Datathon, by using minimal background information. They used a similar methodology to re-identify from this dataset a Member of Parliament using publicly available tweets.

Demonstrating the ease of attacking de-identification methods does not always lead to improvements in de-identification methods and data sharing protocols. Sweeney (2015) showed that numerous individuals in a patient-level health dataset sold by Washington State sans names and

[24] Ryan Singel, Netflix spilled your Brokeback Mountain Secret, Lawsuit Claims, 17 December 2009 https://www.wired.com/2009/12/netflix-privacy-lawsuit/;;
[25] A copy of the lawsuit is available at https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf
[26] The dataset removed all demographic data except the patient's gender and year of birth. Culnane et. al. were able to demonstrate that with an individual's year of birth and some information about the date of a surgery or other medical event, the individual could be re-identified.

addresses, could be reidentified by using a single searchable news repository, thereby putting names to patient records. Washington State responded to the study by requiring a data sharing agreement of the publicly available version of the data and making a more detailed version available only through an application process (Yoo et. al. 2018).[27] Other states also sold similar list but not all states adopted more privacy protecting measures among them were Vermont and Maine. Yoo et. al. (2018) used a similar reidentification strategy to that of the earlier Sweeney (2015) test on the patient-level health dataset from these states. They found that 28.3% of the names from local news stories uniquely matched to one hospitalization in the Maine hospital data and 34% of the names matched to one hospitalization in the Vermont hospital data. Despite the use of Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor standard, the Maine data still allowed eight matches (3.2%) and the Vermont data allowed five matches (10.6%) (Yoo et. al. 2018).

Even datasets with strong anonymization measures are not immune. Abowd et. al. (2023) using only published data, found that an attacker could verify all records in 70% of all census blocks, equivalent to 97 million people for U.S. 2010 census. Abowd et. al. (2023). claim that the use of weaker statistical disclosure framework for tabulations as opposed to microdata for the 2010 census, enabled them to convert these tabulations to microdata.

A survey by the Census Bureau in preparation to the 2020 Census showed 53% of respondent were concerned about the confidentiality of their census responses. Of these concerned respondents, 28% were “extremely concerned” or “very concerned” and a further 25% were “somewhat concerned” about the confidentiality of their census responses (McGeeney et. al. 2019). This survey also showed that 42% of Non-Hispanic Asians and 38% of Non-Hispanic Blacks/African Americans were the most likely to be “extremely concerned” or “very concerned” that the Census Bureau would not keep answers to the 2020 Census confidential. (McGeeney et. al. 2019).

[27] https://doh.wa.gov/data-and-statistical-reports/health-statistics/data-request-faq; Last visited 10 April 2024

Those re-identified in publicly available de-identified data sets have also suffered actual harm. In 2021, a newsletter covering the Catholic Church used 24 months' worth of commercially available records of app signal data covering portions of 2018, 2019, and 2020, which included records of Grindr usage and locations where the app was used, to identify and reveal the sexual orientation of a high ranking member of the clergy, who later resigned.[28] Thus, the concerns of preserving privacy and anonymity in the big data context raised by Heffetz and Ligett (2013), a decade ago, still remain given that there is a risk of more harm due to both the often-sensitive content and the vastly larger numbers of people affected.

4.4 Sharing/selling personal information and the privacy paradox.

In view of the amount of data that can be possibly linked to an individual, it is not surprising that attitudes towards the sale and sharing of personal data to third parties are overwhelmingly negative:

- A 2018 survey conducted on behalf of the Australian Competition and Consumer Commission showed that 86% of respondents consider sharing information with unknown third parties to be a misuse of personal information.[30]
- The 2018 Ipsos Global Advisor survey on attitudes toward data privacy in partnership with the World Economic Forum showed that 62% of consumers feel they should be able to refuse letting companies use personal data and should be paid or rewarded for it. Another 64% of consumers are comfortable sharing personal data with companies if there is a promise not to share them or not to sell them to other parties.[31]

[28] The Pillar, "Pillar Investigates: USCCB gen sec Burrill resigns after sexual misconduct allegations" 21 July 2021, available at https://www.pillarcatholic.com/p/pillar-investigates-usccb-gen-sec
[30] https://www.accc.gov.au/system/files/ACCC%20consumer%20survey%20-%20Consumer%20views%20and%20behaviours%20on%20digital%20platforms%2C%20Roy%20Morgan%20Research.pdf
[31] https://www.ipsos.com/sites/default/files/global-citizens-data-privacy.pdf

- A 2019 survey conducted on behalf of the United Kingdom Information Commissioner's Office (UK ICO) which showed that 67% of respondents found it unacceptable for their details to be sold to or shared with other organizations.[32]
- A 2019 survey by Pew Research showed that 67% respondents say they understand little to nothing about what companies are doing with their personal data. 79% of these respondents claim that they are very or somewhat concerned about how companies are using the data they collect about them. 81% of these respondents also claimed that they have very little or no control over the data collected about them by. [33]
- A 2020 survey by McKinsey resulted in 71% of respondents claiming they would stop doing business with a company if it gave away sensitive data without permission.[34]
- A 2020 survey by Genesys showed that 91% of consumers are concerned that companies might abuse their personal data, with 32% expressing themselves as very concerned.[35]
- A 2020 survey by Ernst and Young indicated that 77% of respondents are opposed to governments selling their personal data to a private sector company. [36]
- A 2022 survey by DataGrail showed that 77% of respondents believe they should be able to opt out of a company selling their data to a third party. Another 79% of respondents expect to have control over how businesses use their data.[37]
- A 2023 survey by Pew Research showed that 67% respondents say they understand little to nothing about what companies are doing with their personal data. 81% of these respondents claim that they are very or somewhat concerned about how companies are

[32]https://ico.org.uk/media/for-organisations/documents/2618466/ico-data-brokers-research-presentation-2903191.pdf
[33]https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/
[34] https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative
[35] https://www.genesys.com/company/newsroom/announcements/122819?release=122819
[36]https://www.ey.com/en_sa/government-public-sector/how-can-digital-government-connect-citizens-without-leaving-the-disconnected-behind
[37] https://www.datagrail.io/resources/pdfs/privacy-and-ecommerce-report/

> using the data they collect about them. 77% of these respondents also claimed to have little or no trust that companies will admit mistakes and take responsibility when they misuse or compromise data. 76% of these respondents also claimed to have little or no trust that social media companies will not sell users' personal data to others without their consent. 84% were very worried or somewhat worried that companies selling your information to others without you knowing. [38]

Notwithstanding the results of these surveys, there is a lot of data that can be purchased often at a low cost. Sherman et. al. (2021) found that U.S. data brokerage industry gathers data on virtually every American, by scraping public records, embedding code into mobile apps, and purchasing customers data from various companies. Sometimes, companies purchase data directly from each other as shown in the recent Federal Trade Commission (FTC) filings against various data brokers such as Kochava.[40]

Even companies that claim not to sell personal data are reported to engage in the bartering of said data for a valuable consideration such as increased engagement on a platform.[41] This distinction between selling and bartering would probably not assuage the concerns of many about the spread and use of their data by companies with whom there is no contract or even a contact. Such is the scope of the data gathered by these brokers that Muralidhar & Palk (2020) claim that it is harder it is to request such data from the government than it is to purchase detailed data about a population from data brokers. The collection of excessive amounts of personal information from users for so-called secondary uses such as selling the data as a commodity, has been well documented by researchers (Spiekermann et. al. 2015; Muralidhar & Palk 2020; Schauer & Schnurr 2023; and Ichihashi 2020).

---

[38] https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/

[40] https://www.ftc.gov/legal-library/browse/cases-proceedings/ftc-v-kochava-inc; The amended complaint against Kochava illustrates what can be described as dark patterns however the FTC uses the broader term "unfair or deceptive acts or practices in or affecting commerce".

[41] See https://apiacademy.co/2018/06/how-the-facebook-api-led-to-the-cambridge-analytica-fiasco/; See also https://www.dli.tech.cornell.edu/post/facebook-and-google-are-the-new-data-brokers; See also

The prevalence of data collected despite an individual's privacy concerns can be explained by work of Athey et. al. (2018) on digital privacy paradox. Their research indicates that even small monetary incentives this leads individuals to surrendering their personal information and drives platforms to gather increased information. Athey et. al. (2018) further posit that this may undermine the efficacy of regulation. The research by Acemoglu et. al. (2022) demonstrates that there is excessive data sharing since everyone will overlook privacy concerns and surrender their own information because others' sharing decisions will have already revealed much about them. They identified a negative externality wherein one comprises privacy not only of oneself but also others, when sharing personal information. Another possible explanation is provided by Acquisti et. al. (2016), such that there may not be a privacy paradox, rather attitudes may not correlate or predict behaviors. As they pointed out, attitudes are often expressed generically but behaviors are specific and contextual.

The economic framework proposed by Jones & Tonetti (2020) to study data shows that when firms own data, they may overuse it and not adequately respect consumer privacy. Acquisti et. al. (2016) described privacy as a decision on what a person wants to protect and what she wants to share at any given moment and in any given context. This brings us back to dark patterns since many are deployed to manipulate individuals into surrendering their privacy. Thus, individuals are unable to accurately price their personal information in relation to the cost of surrendering it despite the prevalent negative attitudes towards this surrender of personal information. Acquisti et. al. (2016) theorizes that is due to information asymmetries such as individuals failing to comprehend the extent to which their personal information is collected and identified online. Spiekermann et. al. (2015) even speculated that knowledge of the volume and business done with their data among third parties, may lead individuals to be surprised and feel betrayed.

4.5 Estimating Reidentification Risks of Dark Patterns in Philippine Survey Data

In this section, we evaluate the disclosure risk of Philippine datasets using two primary measures: the Risk of Reidentification (RRI) and k-anonymity. Additionally, we analyze the interconnection and susceptibility of publicly available datasets by linking Personally Identifiable Information (PII) records.

Microdata from following surveys was examined:

1. 2022 Philippine National Demographic and Health Survey (NDHS) by United States Agency for International Development Aid (USAID);
2. 2022 Annual Poverty Indicators Survey (APIS) by the Philippine Statistics Authority (APIS); &
3. Consumer Expectations Survey (CES) by the Bangko Sentral ng Pilipinas.

The Risk of Re-identification or Reidentification Risk refers to the potential that supposedly anonymous or pseudonymous datasets could be de-anonymized to recover the identities of users.[42] To compute the RRI, all entries in a dataset will be compared with other entries. Entries with the same values for indirect identifier variables (i.e., variables that do not directly associate with a specific individual) will form an 'equivalency class.' For instance, entries in a dataset with similar gender, age, and city address, like 35-year-old men residing in Manila, would form an equivalence class. The size of an equivalence class is equal to the number of entries whose indirect identifiers have the same values.

Larger equivalence classes have lower probabilities of re-identification as they have more data subjects to deal with. Conversely, smaller equivalence classes have higher probabilities of re-identification because these have fewer data subjects with the same indirect identifiers.

Once the risk for each row is known using the below equation, the RRI of the whole dataset can be calculated by getting the mean of all per-row RRIs.

[42] Commissaire à l'Information et à la Protection de la vie Privée (2016). "De-identification Guidelines for Structured Data". Available at https://www.ipc.on.ca/p-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

$$Risk\ of\ Re-identification = \frac{1}{Size\ of\ Equivalence\ Class}$$

Additionally, sensitive rows violating the k-anonymity condition are consolidated to gauge the level of disclosure risk. k-anonymity is a data anonymization technique used for reducing the risk associated with releasing individual-level data by ensuring that each entry in the dataset is indistinguishable from at least k-1 other entries, based on a set of quasi-identifiers. An entry violates k-anonymity if the frequency (f) of entries with a certain combination of PIIs is smaller than the specified threshold (k). In simpler terms, if there are insufficient entries in the dataset with identical quasi-identifiers, it fails to satisfy the k-anonymity requirement, making it susceptible to disclosure risk. For instance, if only two records share a unique combination of quasi-identifiers not found in any other record in the dataset, it does not meet the standard of 3-anonymity.

The RRI and count of sensitive rows of the three sample datasets are computed and listed in Table 1. Masked data were also generated using data masking techniques, such as generalization and suppression, and underwent RRI estimation.

Table 1: Description of the three sample datasets

	No. of columns	**No. of rows**	**No. of PIIs**	**% of Sensitive Rows (k=3)**
NDHS	619	129,724	53	18.79%
APIS	75	179,947	9	25.11%
CES	1,361	7,468	26	5.58%

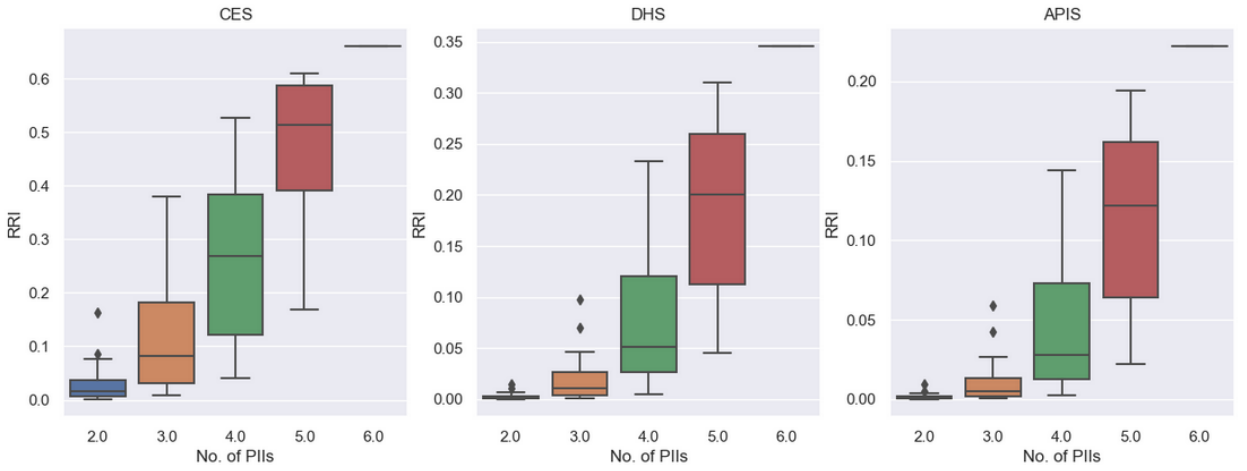
Table 2: Risk Re-Identification Indices (RRI) of the three sample datasets

	No Masking	**With Data Masking**
NDHS	0.3459	0.0635
APIS	0.2223	0.1149
CES	0.6613	0.2248

Table 1 shows the RRI for the three (3) sample datasets using seven (7) common Personally Identifiable Information (PII) variables.[43,44] We applied data masking to the three datasets by binning and generalizing the PIIs and compared the RRI before and after this procedure. For all datasets, it can be observed that the risk is lower for masked datasets, which increases the size of the equivalency classes. Furthermore, the CES yielded the highest RRI values compared to NDHS and APIS. This could be attributed to the CES having the smallest number of rows, which limits the size of the group of individuals sharing the same attributes and thus increases the risk of reidentification.

We note that the disclosure of PIIs correlates with an increase in the RRI. As the number of disclosed PIIs increases, it becomes easier to link and determine individuals from different datasets (see Ochoa et al, 2001; Torra & Navarro-Arribas, 2023). To demonstrate this, we generated datasets, each comprising different combinations and numbers of PIIs, and calculated the RRI. As can be observed from the boxplots below, on average, the RRI increases as the number of disclosed PIIs increases. However, masking the datasets significantly reduces the RRI as evident in Figure 2.

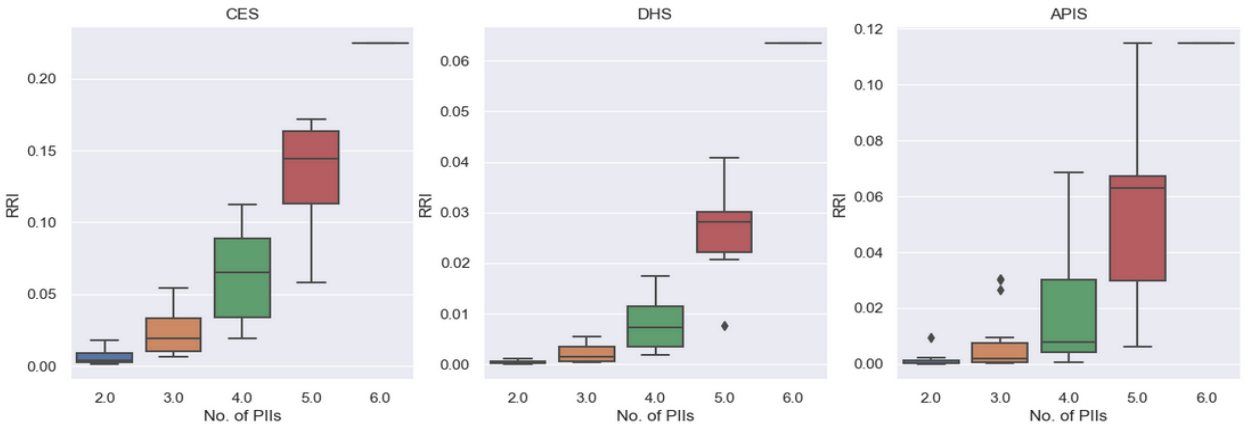
Figure 1. Boxplots of RRI for unmasked datasets using different numbers of PIIs.



[43] These variables are region, age, sex, relationship to household head, marital status, highest educational attainment, and number of household members.

[44] with exception to the APIS dataset that only used the first six PIIs mentioned above.

Figure 2. Boxplots of RRI for masked datasets using different numbers of PIIs



In addition to the RRI, we also calculated the percentage of sensitive rows within the 3 datasets, considering different values of *k*.[45] In the context of *k*-anonymity, *k* refers to a parameter that determines the level of anonymity provided to individuals in a dataset. An increase in *k* indicates stronger privacy protection in terms of re-identifying risks.

Figure 3: *k* vs Percentage of Sensitive Rows

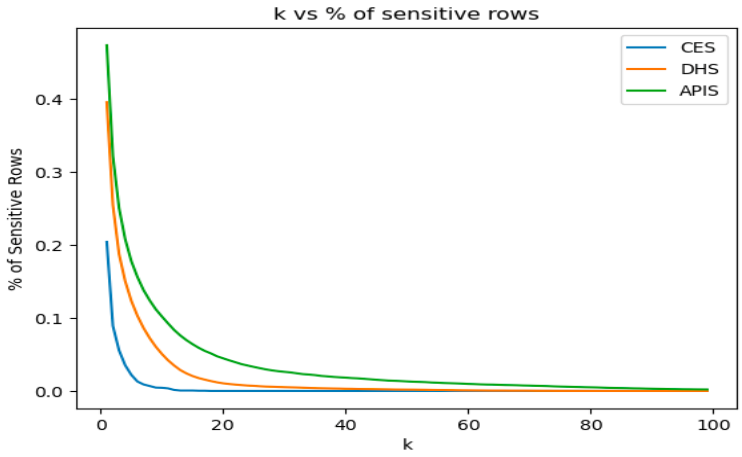


Figure 3 shows that, for all datasets, the percentage of sensitive rows decreases as *k* increases. This phenomenon occurs because *k*-anonymity requires that each equivalence class within the dataset

[45] We considered $k \in \{1, 2, 3, \ldots, 100\}$.

contains at least *k*-indistinguishable records. Consequently, more records must be clustered together within each equivalence class to meet this criterion, which in turn results in a more generalized dataset and decreases the count of unique sensitive rows.

As demonstrated, data masking effectively mitigates disclosure risk by obfuscating or replacing sensitive data. Nevertheless, it is essential to recognize that masking methods can reduce several risks but cannot eliminate them, especially when facing intentional and sophisticated attacks. No system is invulnerable to determined attackers. Thus, it is crucial to implement additional security measures and remain vigilant to minimize the impact of potential breaches.

With the proliferation of publicly available datasets and the emergence of unconventional access methods such as dark patterns, there arises the capability to interconnect multiple datasets to identify potential overlaps. For instance, given our sample data, we used variables such as region, age, marital status, relationship to head, and highest educational attainment, as mapped in the figure below, to connect the three datasets together.

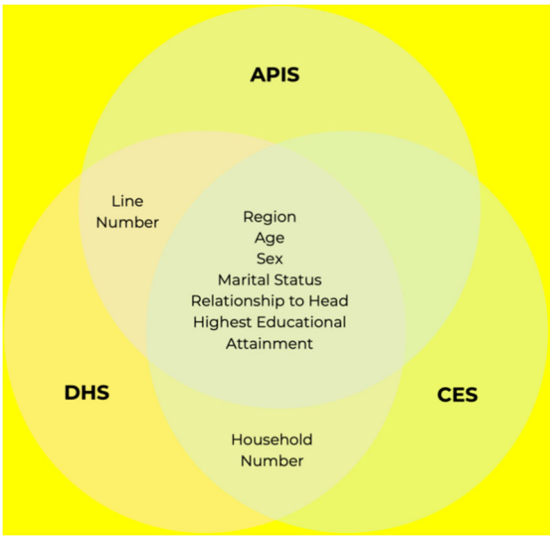


Figure 4: Venn Diagram of PIIs in the APIS, DHS, and CES datasets

Using the sample datasets, the deterministic linkage technique (Rocha et. Al. 2019, Pereira et. al. 2023) is employed to explore the possibility of association. This involves systematically identifying and associating corresponding records across different datasets based on shared identifiers or characteristics. From a privacy perspective, this linkage raises significant concerns as it aims to find exact matches or highly probable matches between records.

The figure presented below shows that 2,038 unique combinations of PIIs were shared by the 3 sample datasets. These combinations represent instances where various pieces of sensitive personal data intersect, which could potentially expose individuals to disclosure risk.

This finding has significant implications for data privacy, especially because these datasets are publicly available and thus more vulnerable to disclosure risk. This highlights the importance of implementing appropriate data anonymization and protection measures to mitigate the risk of privacy breaches.

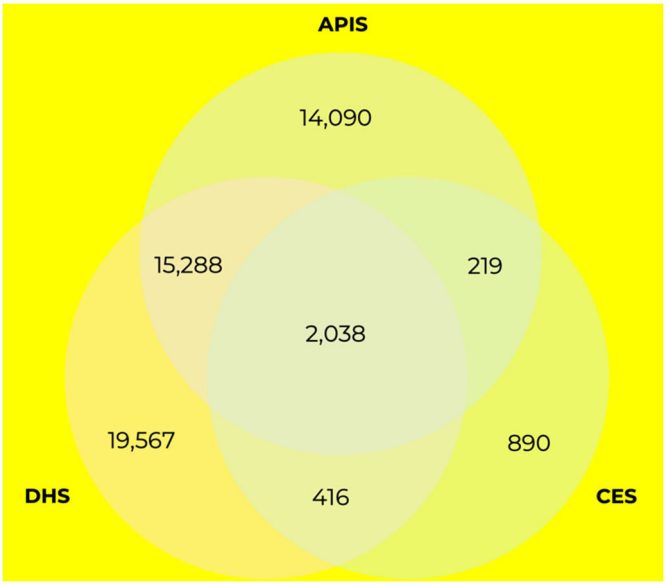


Figure 5: Matched Records in Sample Datasets using Deterministic Linkage

Linkage attacks on this information can be easily performed by purchasing or securing datasets from numerous brokers. As hinted at by numerous researchers, it is likely that the public is unaware of the amount of their personal data that is being openly sold (Spiekermann et. al. 2015; and Sherman et. al. 2023).

Data brokers such as DB to Data, a Philippine based data broker openly sell call and email lists online.[46] DB to Data offers Philippine call lists with prices ranging from USD 150.00 for a list of 10,000.00 to USD 4,000.00 for a list of 3,000,000.00. The lists include the names, cellphone numbers, address, relationship status, age and work. Lists are also openly available for other countries including the Germany (USD 6,000.00 for a list of 6,000.000.00) and United States (USD 50,000.00 for a list of 90,000,000.00). Special categories such as Binance (USD 4,000.00 for a list of 3,000,000.00) and Chinese Overseas British (USD 6,000.00 for a list of 100,000.00) are also available. There does not seem to be any restrictions on the purchase of the data. There is also no ready explanation as to the source of this data except that it has been verified by both a human and a computer. DB to Data is by no means unique. Buymailmarketinglists.com,[47]. AllEmailList,[48] DBShop,[49]and probably other companies offer comparable products with a similar lack of restrictions on purchase.

The prices and variety of the microdata available from these presumably legitimate companies is dwarfed by the microdata that is available for sale on social media networks. Facebook Groups such as "USA, UK, AUS, CAN, Email Data Buy Sell for Tech Support"[50], "High Ticket Closers and Setters (JOB OPPORTUNITIES),[51] "E-mail Blasting Data-Buyers and Sellers"[52] are among the numerous venues where

[46] Per the Philippine National Privacy Commission, there is no record of DB to Data being registered as a personal information controller which is required by Philippine law for entities processing the sensitive personal information of more than 1,000 individuals.
[47] This is a Sri Lanka based broker, it offers a list of over 500,000 individuals in the Philippines for USD 89. See https://www.buyemailmarketinglists.com/
[48] The Philippine database for sale by AllEmailList is alleged to be update monthly and contains no usage limitations. See https://www.allemaillist.com/philippines-cell-mobile-phone-number-list-database.html
[49] Includes Philippine data per province. No prices indicated. See https://dbshop.in/database/states/philippines/173
[50] Over 2,100 members, see https://www.facebook.com/share/hpNRfAz5P7AxPjQ9/?mibextid=A7sQZp
[51] Over 42,000 members, see https://www.facebook.com/share/FePtfvt2kend4dzW/?mibextid=A7sQZp
[52] Over 4,700 members, see https://www.facebook.com/share/qen3HbUKdzGRiPCa/?mibextid=A7sQZp

the sale of microdata can take place. One online seller in these groups offered email lists with 450,000,000 entries on individuals based in the United States for the price of USD 1,500.00 and email lists of 10,000,000 Filipinos for USD 400.00.[53] The number of sellers in these groups dwarfs the 540 data brokers that are registered in both or either California and Vermont, as of 2021, as required by their respective state laws.[54] Data brokers appearing in Google searches and Facebook are only 2 of the very many possible channels for which individual data is bought and sold. These are just the small-fry, in 2020, Axciom, one of the biggest data brokers, declared that is has "*the largest depth and breadth of demographic segmentation and predictive data, identifying over 260 million US consumer portraits*.".[55] Axciom claims that it has the *"largest catalog of over 12,000 global data attributes specifically focused on providing personalized experiences."*[56] Axciom's data is sourced among others from "*other commercial entities where consumers have been provided notice of how their data will be used, and offered a choice about whether or not to allow those uses*."[57]

Simply put, for the price of a few dollars, one can possibly re-identify a person out of a government released data set. For the price of a few hundred dollars, one can possibly link all the information in said dataset with additional information, producing a very in-depth profile of all individuals.

5. The Ethical and Legal Frameworks

Hasan et. al (2020) have pointed out that privacy and protection of data is one the biggest critical issue of big data services. Many institutions have rich repositories of data but not all have the necessary expertise or care that would allow the safe and ethical sharing of data. In addition to any legal obligations to do so, many researcher routinely promise anonymity to subjects who participate in studies or surveys (Heffetz & Ligett). These promise of anonymity hold immense value to respondents, as a survey by Hotz

[53] Person will not be named due to possible privacy implications. However, from the profile, it is unclear if this is an individual or a group working together as a data brokerage.
[54] https://privacyrights.org/resources/registered-data-brokers-united-states-2021
[55] https://business.linkedin.com/content/dam/me/business/en-us/amp/marketing-solutions/images/lms-partner/partner-dedicated-acxiom/pdf/top-data-and-models-for-digital-by-industry-vertical-flyer.pdf
[56] https://www.acxiom.com/customer-data/
[57] https://www.acxiom.com/wp-content/uploads/2013/09/Acxiom-Marketing-Products.pdf

and Slanchev (2017) showed that 79.5% of respondents cited confidence in researchers to keep responses and information private" as the most important determinant for their participation in a hypothetical study.

As a recognition of disclosure risks, datasets may be so extensively altered to limit disclosure risks (e.g. noise added), subject to onerous conditions for release or simply data may not be released at all (Muralidhar & Palk 2020; Hotz et. al. 2022; and Groshen & Goroff 2023). These conditions or limitations, constrict the spread of the data and presumably prevent valuable innovations or insights from occurring, thus begging the question, why did we bother collecting the data at all, if we are not going to use it for societal progress (Oberski & Kreuter 2020). Various countries have passed laws which try to balance these often competing needs for utility of the data with the need for privacy of the individuals from whom the data was sourced. Thus, a close look at the interaction between current and proposed regulations on dark patterns and disclosure risks is in order.

5.1 Minimization of Disclosure Risks

The United Nations General Assembly adopted the Fundamental Principles of Official Statistics in January 2014.[59] Principle 6 of which states: *"Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."* The adoption and implementation of these principles is in various national laws, overviews of which can found on the United Nations Statistical Division website.[60]

Going beyond statistical data and into rich administrative datasets held by government, Muralidhar & Palk (2020) documented the history and scope of privacy obligations of governmental privacy obligations on researchers and the public in the United States. They detail the scope of freedom

[59] https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf
[60] https://unstats.un.org/unsd/dnss/cp/searchcp.aspx

of information requests for research data in contrast with obligations to keep data confidential such as the Privacy Act of 1974 and the HIPAA.

For the sake of illustration, a researcher may be granted access to confidential data held by the U.S. Census Bureau, if it is needed to undertake a task that will contribute substantially to Census Bureau programs and only if the data can be adequately protected. These researchers may be granted special sworn status under Title 13, Section 23 of the U.S. Code, which makes them subject to the same legal obligations and penalties as regular Census Bureau staff. The process for access can take a maximum of 12 weeks. The output of such research will be subject to the disclosure avoidance review process to ensures that project output does not reveal confidential information about individual people, households, or firms, which can take another 6 weeks.[61] Ruggles et. al. (2018) critiques the process for this confidential access stating that among others, that most research topics ineligible since they are approved only if they benefit the Census Bureau, and the process is time-consuming and costlier than using public use data.

Similarly, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) enacted in 2002 as Title V of the E-Government Act of 2002, contains similar concepts to Title 13 in order to provide federal agencies with data confidentiality protections.[62] The Foundations for Evidence-Based Policymaking Act of 2018 ("Evidence Act") reauthorized the CIPSEA as Title III of the former law, with a goal of improving the process for researchers to access government data by making it more open, streamlined, and secure, while also providing stronger privacy standards and legal protections. [63] One of the highlights is CIPSEA pledge, wherein information acquired under the CIPSEA pledge: a. is used exclusively for statistical purposes; b. cannot be disclosed in identifiable form to anyone not authorized by CIPSEA; and C. is safeguarded by control of its access and use. [64]

[61] https://www.census.gov/topics/research/guidance/restricted-use-microdata/standard-application-process.html
[62]https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/inforeg/proposed_cispea_guidance.pdf
[63]https://uscode.house.gov/view.xhtml?path=%2Fprelim%40title44%2Fchapter35%2Fsubchapter3&edition=prelim
[64] See Bureau of Labor Statistics CIPSEA report for 2022; available at https://www.bls.gov/bls/cipsea-report.htm

The European Union (EU) also imposes a legal obligation on European Statistical System members to protect confidential data as contained in Chapter V of Commission Regulation (EC) No 223/2009 on European statistics.[65] The procedure for accessing microdata in the EU is like that of the US Census Bureau. EU Regulation No 557/2013 is the legal basis for scientific access to European microdata.[66] Among other restrictions and procedure, it provides for access within access facilities at Eurostat or accredited by Eurostat.[67] In a sense, it is more restrictive as Eurostat also requires that the researcher be part of recognized research entity, among other requirements.[68]

Similar to the US Evidence Act, the EU has also implemented the European Data Governance Act (EU DGA).[69] The EU DGA is a key pillar of the European Strategy for Data.[70] A key feature of the EU DGA is the concept of data altruism, which is intended to allow easier sharing data based on their consent for purposes of general interest by individuals and businesses.[71] The EU DGA also introduced the requirements for registration and operation of data altruism organisations which will allow collection by data for general interest purposes subject to requirements meant to ensure trust and confidence in these organisations.[72] Nonetheless, public sector bodies may impose requirements for the re-use of protected data such as anonymization for personal information, [73] disclosure controls such as aggregation, or access

[65] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32009R0223, last visited 29 March 2024
[66] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0557
[67] Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002 Text with EEA relevance; available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0557; last visited 29 March 2024
[68] https://ec.europa.eu/eurostat/documents/203647/771732/Recognised-research-entities.pdf; For an example on how these recognized entities access Eurostat Micro data see Comisión Económica para América Latina y el Caribe (CEPAL) guide available at https://biblioguias.cepal.org/eurostat/about
[69] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R0868
[70] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
[71] Article 2, Paragraph 16 of the EU DGA defines Data Altruism as "*'data altruism' means the voluntary sharing of data on the basis of the consent of data subjects to process personal data pertaining to them, or permissions of data holders to allow the use of their non-personal data without seeking or receiving a reward that goes beyond compensation related to the costs that they incur where they make their data available for objectives of general interest as provided for in national law, where applicable, such as healthcare, combating climate change, improving mobility, facilitating the development, production and dissemination of official statistics, improving the provision of public services, public policy making or scientific research purposes in the general interest;*"
[72] Chapter IV of EU DGA covers data altruism organisations.
[73] In the EU anonymized information is information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. See GDPR Recital 26 available at https://www.privacy-regulation.eu/en/recital-26-GDPR.htm

within secure environments.[74] Data altruism organizations will be monitored and supervised by the authorities competent for the registration of altruism organizations, established by each member EU state.[75]

These obligations to keep personal information confidential run in parallel to the various laws to protect personal information and privacy rights. Oberski & Kreuter (2020) enumerated several examples of legislative measures to increase the protection of personal information such as the 2018 European General Data Protection Regulation (GDPR), the 2017 Japanese Amended Act on the Protection of Personal Information the 2020 Brazilian General Data Protection Law and the 2020 California Consumer Privacy Act (CCPA), among others. The list is growing with Colorado, Connecticut, Oregon, Texas, Utah and Virginia, enacting data privacy laws that became effective in 2023 & 2024. Several other states will have data privacy laws taking effect later in 2024 till 2026. Moreover, United States which already has an arguably robust patchwork for the protection of personal information, is looking into enacting a comprehensive Federal law on data protection or the American Privacy Rights Act.[76]

Philips et. al. (2017) notes that the recognition of the likelihood of reidentification, due in part to big data, has raised calls for the criminal penalties for unauthorized re-identification of anonymized information. At least two jurisdictions that impose criminal penalties on the re-identification of anonymized information. Section 48F of the Singapore's Personal Data Protection Act 2012 criminalizes the unauthorized re-identification of anonymized information.[77] Similarly, Section 171 of the United Kingdom's Data Protection Act 2018 makes it an offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data.[78]

---

[74] Article 5 of the EU DGA.
[75] See Article 17 of the EU DGA. A list of data altruism organisations will be made available at https://digital-strategy.ec.europa.eu/en/policies/data-altruism-organisations
[76] https://iapp.org/news/a/new-draft-bipartisan-us-federal-privacy-bill-unveiled/
[77] https://sso.agc.gov.sg/Act/PDPA2012%3FViewType%3DPdf%26_%3D20210111164941&ved=2ahUKEwjet-3N3tWFAxXT8qACHTNkCNUQFnoECB8QAQ&usg=AOvVaw1i7-UgUzI37xD8Rh_fi5_R
[78] https://www.legislation.gov.uk/ukpga/2018/12/section/171?view=plain

5.2 Regulation of Dark Patterns

Dark patterns have been recognized as a distinct concept for over a decade, but it is only recently that legislation has been passed to regulate them. The Organisation for Economic Co-operation and Development (OECD) in its paper on dark commercial patterns (OECD 2022) doubts if market forces can address dark patterns alone and points out that at times the markets incentivize their use. In the U.S.A., California, Colorado, and Connecticut, have enacted legislation to regulate or ban certain forms of dark patterns. Section 1798.140.l of California Consumer Privacy Act (CCPA) defines dark patterns as a "user interface designed or manipulated with the substantial effect of subverting or impairing user autonomy, decision making, or choice, as further defined by regulation." Section 1798.140.h of CCPA notably states that "*agreement obtained through use of dark patterns does not constitute consent*." The Connecticut Data Privacy Act,[79] the Colorado Privacy Act,[80] and the Vermont Data Privacy Act[81] are similar state laws that also regulate dark patterns.

California also enacted the California Age-Appropriate Design Code Act (CAADCA), to take effect on 01 July 2024.[82] Among other provisions aimed at protecting children, the CAADCA prohibits a business that provides an online service, product, or feature likely to be accessed by children from using dark patterns to lead or encourage children to provide personal information beyond what is reasonably expected to provide that online service, product, or feature to forego privacy protections, or to take any action that the business knows, or has reason to know, is materially detrimental to the child's physical health, mental health, or well-being. However, NetChoice, LLC, a national trade association with members from the tech and social media industry obtained a preliminary injunction from the District Court for the Northern District of California preventing the State of California from enforcing the CAADCA.[83] NetChoice,

[79] https://www.cga.ct.gov/2022/act/Pa/pdf/2022PA-00015-R00SB-00006-PA.PDF
[80] https://leg.colorado.gov/sites/default/files/2021a_190_signed.pdf
[81] https://legislature.vermont.gov/Documents/2024/Docs/BILLS/H-0121/H-0121%20As%20passed%20by%20the%20House%20Official.pdf
[82] https://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202120220AB2273&showamends=false
[83] https://netchoice.org/wp-content/uploads/2023/09/NETCHOICE-v-BONTA-PRELIMINARY-INJUNCTION-GRANTED.pdf

LLC argues that the CAADCA violates the First Amendment since it: (1) is an unlawful prior restraint; (2) is unconstitutionally overbroad; (3) is void for vagueness; and (4) is subject to and fails strict scrutiny.[84] The case is still pending.[85]

Prior to specific legislation enacted to combat dark patterns, the FTC has used the Federal Trade Commission Act, 15 U.S.C. § 57a(a)(1)(b) prohibiting unfair or deceptive trade practices, as the basis for penalizing entities that make use of dark patterns.[86] The FTC's September 2022 report, *Bringing Dark Patterns to Light*, details many of these cases.[87]

In the EU, dark patterns are not specifically mentioned in the GDPR. However they may still violate the fairness and transparency principle in article 5(1)(a), the accountability principle in article.5(2), data protection by design and default in article 25, the requirement to provide transparent privacy notices to data subjects in articles 12(1), 13 & 14), and the data subject rights in articles 15 to 22 of the GDPR. [88] Specifically, on 14 February 2023, the European Data Protection Board ("EDPB") adopted Guidelines 03/2022 on Deceptive Design Patterns in Social Media Platform Interfaces: How to Recognise and Avoid Them.[89] For version 2.0 of these Guidelines 03/2022, the EDPB decided on using the more inclusive and descriptive term "deceptive design pattern" instead of "dark pattern". Guidelines 03/2022, defines deceptive design pattern as "*interfaces and user experiences implemented on social media platforms that lead users into making unintended, unwilling and potentially harmful decisions in regards to their personal data with the aim of influencing users' behaviors*". It also defines six (6) categories of dark patterns; (1) overloading, (2) skipping, (3) stirring, (4) hindering, (5) fickle, and (6) left in the dark. The EDPB Guidelines

[84] https://netchoice.org/wp-content/uploads/2022/12/NetChoice-v-Bonta_-Official-AB-2273-Complaint-final.pdf
[85] https://oag.ca.gov/system/files/attachments/press-docs/NetChoice%20Ninth%20Cir.%20Opening%20Brief.pdf
[86] http://uscode.house.gov/view.xhtml?req=granuleid%3AUSC-prelim-title15-chapter2-subchapter1&edition=prelim
[87] https://www.ftc.gov/reports/bringing-dark-patterns-light
[88] https://media.squirepattonboggs.com/pdf/Data-Protection/Uncloaking_Dark_Patterns_Identifying_Avoiding_And_Minimizing_Legal_Risk.pdf
[89] Said Guidelines were released for public consultation on 14 March 2022. Subsequently, the EDPB adopted version 2.0 of these Guidelines on 14 February 2023, which is available at https://www.edpb.europa.eu/system/files/2023-02/edpb_03 2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf; last visted on 27 March 2024.

03/2022 only provide recommendations and guidance for the design of the interfaces of social media platforms, thus limiting both its scope and enforceability.

In 2023, the UK ICO wrote to 53 of the UK's top 100 websites to warn of enforcement action if the latter do not make changes to their cookie banners in order to comply with the Privacy and Electronic Communications Regulations 2003 (as amended) (PECR), the Data Protection Act 2018 (DPA) and the UK General Data Protection Regulation 2018 (UK GDPR).[90] The UK ICO reported a positive reception to their warnings as 38 of the 53 organisations have changed their cookies banners to be compliant and 4 have committed to reach compliance before the end of February 2024. The UK ICO relayed that it planned to write to the next 100 websites and the 100 after that.[91]

The recent EU Digital Services Act (EU DSA) which recently came into effect last February 17, 2024, adds to the regulatory framework for dark patterns.[92] It defined dark patterns on online interfaces of online platforms as "*practices that materially distort or impair, either on purpose or in effect, the ability of recipients of the service to make autonomous and informed choices or decisions. Those practices can be used to persuade the recipients of the service to engage in unwanted behaviours or into undesired decisions which have negative consequences for them. xxx xxx xxx*"[93] The EU DSA covers all intermediary services that offer their services to users based in the EU, including online platforms such as app stores, collaborative economy platforms, and social media platforms.[94] Article 25 of the EU DSA stipulates that providers must not design online platforms in a deceitful manner that would impair recipients' ability to make free and informed decisions such as promoting certain user choices over others; repeated requests

[90] https://ico.org.uk/media/about-the-ico/documents/4027811/cookie-banner-concerns.pdf
[91] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/01/ico-warns-organisations-to-proactively-make-advertising-cookies-compliant/
[92] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065&qid=1692301215242
[93] See paragraph 67
[94]See Article III (g) of the DSA which states "'intermediary service' means one of the following information society services:
(i) a 'mere conduit' service, consisting of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network;
(ii) a 'caching' service, consisting of the transmission in a communication network of information provided by a recipient of the service, involving the automatic, intermediate and temporary storage of that information, performed for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request;
(iii) a 'hosting' service, consisting of the storage of information provided by, and at the request of, a recipient of the Service"

to the recipient to make a choice which has already been made; and termination of the service being made more difficult than initial subscription or sign-up.

The EU DSA imposes stricter obligations on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), given the structural and "systemic" significance of certain firms in the digital services ecosystem.[95] Per Article 33 of the EU DSA, VLOPs and VLOSEs have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million, and which are designated as such by the European Commission.[96] VLOPs and VLOSEs will have to higher transparency standards, provide access to (personal) data to competent authorities and researchers, and identify, analyze, assess, and mitigate systemic risks linked to their services, including an independent audit per Article 37 of the EU DSA.

EU regulators have already been acting even prior to the implementations of the EU DSA and the EDPB Guidelines on Dark Patterns. In C-252/21 Meta Platforms Ireland and Others v. Bundeskartellamt, the Court of Justice of the EU found that to sign up for Facebook, users have to consent to a user-agreement that allows the Facebook to track their movements across the web by using off-Facebook data or online data outside Facebook. The CJEU also found that there is no mechanism to separately consent for the processing of data within the Facebook social network and for the off-Facebook data. Due to the absence of these separate consent mechanisms, the consent of users to the processing of off-Facebook data must be presumed not to be freely given. This is a form of bundled consent, and while the CJEU did not specifically mention dark patterns or deceptive design patterns in its decision, regulators in the UK have labelled it as a harmful online choice architecture practice.[97]

---

[95] See EU DSA recital 53

[96] For an overview of these VLOPs and VLOSEs see https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses; The initial list of VLOPs and VLOSEs is available at https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413

[97] For the purposes of this paper, harmful online architecture is practically synonymous with a dark pattern. In their joint position paper, both the UK Information Commissioner's Office, and Competition & Markets Authority state that: "*Poor OCA practices can manipulate and influence users of digital services to make choices about their personal information that do not align with their preferences, such as sharing more personal information than they would otherwise volunteer.*"; See https://www.drcf.org.uk/__data/assets/pdf_file/0024/266226/Harmful-Design-in-Digital-Markets-ICO-CMA-joint-position-paper.pdf

The France's *Commission Nationale de l'Informatique et des Libertés* (CNIL) imposed a fine of €60,000,000 on Facebook[98] and €150,000,000 on Google for making it more difficult for internet users to refuse cookies than to accept them.[99] The CNIL found that in order to guarantee that consent is freely given, it should be as easy to refuse cookies as to accept them and it in fact cited the study by Nouwens et. al. (2020) as corroborating this conclusion. As a result of this action by the CNIL, Google installed on google.fr and youtube.com a refusal button entitled "Only allow essential cookies" near the acceptance button.

Tiktok was also fined by the CNIL for implementing advertising identifiers sans consent and for insufficiently informative cookie banners.[100] CNIL's authority to investigate TikTok's cookie practices pursuant to Article 82 of the French Data Protection Act which adopted the EU e-Privacy Directive. CNIL penalized Voodoo for still processing despite user refusal, information linked to the browsing habits for advertising purposes in contradiction with what it indicates in the information screen it displays.[101] Tiktok was also fined €345,000,000 and reprimanded by the Ireland's Data Protection Commission (IDPC) for violations of multiple GDPR articles.[102] The IDPC also found that TikTok used dark patterns such as Preselection, Visual Interference, and Forced Action.

While, most extensively studied in the United States and the Europe Union, regulators are noticing dark patterns worldwide. India has recently enacted guidelines prohibiting dark patterns.[103] China has enacted the Personal Information Protection Law of the People's Republic of China[104] which prohibits big data swindling among other dark patterns. [105] The Australian Competition and Consumer Commission

---

[98]https://www.cnil.fr/sites/cnil/files/atoms/files/deliberation_of_the_restricted_committee_no._san-2021-024_of_31_december_2021_concerning_facebook_ireland_limited.pdf

[99] https://www.cnil.fr/en/closure-injunction-issued-against-google

[100] https://www.cnil.fr/en/cookies-cnil-fines-tiktok-5-million-euros

[101] https://www.cnil.fr/en/mobile-games-closure-injunction-issued-against-voodoo

[102] https://www.edpb.europa.eu/system/files/2023-09/final_decision_tiktok_in-21-9-1_-_redacted_8_september_2023.pdf

[103] Guidelines for Prevention and Regulation of Dark Patterns, 2023; available at https://consumeraffairs.nic.in/sites/default/files/fileuploads/latestnews/Draft%20Guidelines%20for%20Prevention%20and%20Regulation%20of%20Dark%20Patterns%202023.pdf; last visited 03 April 2024

[104] Translation available at https://www.china-briefing.com/news/the-prc-personal-information-protection-law-final-a-full-translation/; last visited 03 April 2024

[105] See https://chinamediaproject.org/the_ccp_dictionary/big-data-swindling/; last visited 03 April 2024

issued the Digital Platforms Inquiry which tackled dark patterns among other issues.[106] Even the Philippine National Privacy Commission has issued guidelines on deceptive design patterns.[107]

Despite the more comprehensive scope of EU law as compared with the United States, complaints by the FTC best illustrate how dark patterns increase disclosure risks in a relatively straight forward manner. An application or website uses dark patterns such as deceptive omissions or deceptive presentations to manipulate its users into sharing a large amount of personal information, this personal information is then sold or shared to data brokers. As the means used to gather this information was a dark pattern, the manipulated individual often has little or no idea that they were a victim of this manipulation, and their data is being spread far and wide.

The first notable example is the FTC complaint against GoldenShores Technologies, LLC, and its owner, Erik Geidl.[108] The FTC accused GoldenShores of deceiving users of its flashlight application by surreptitiously collecting personal data including locations and sharing it with third parties. The FTC further accused GoldenShores of not allowing users to disable this collection and sharing of information. In its report on this case, Forbes quoted Jessica Rich, FTC's Director of the Bureau of Consumer Protection as saying: "*But this flashlight app left them in the dark about how their information was going to be used*." The FTC and GoldenShores reached a settlement wherein GoldenShores and Geidl were prohibited from misrepresenting how consumers' information is collected and shared and how much control consumers have over the way their information is used. GoldenShores and Geidl were also required to delete any personal information collected from consumers through the Brightest Flashlight application.[109]

In the second example the FTC accused PaymentsMD of using deceptive omission and deceptive presentation in order for consumers to consent to the collection of sensitive health information from third

[106] https://www.accc.gov.au/system/files/Digital%20platform%20services%20inquiry.pdf
[107] https://privacy.gov.ph/wp-content/uploads/2023/11/NPC-Advisory-No.-2023-01-Guidelines-on-Deceptive-Design-Patterns_7Nov23.pdf
[108] https://www.ftc.gov/sites/default/files/documents/cases/131205goldenshorescmpt.pdf
[109] https://www.ftc.gov/system/files/documents/cases/140409goldenshoresdo.pdf

parties.[110] PaymentsMD launched a fee-based service called Patient Health Report, that would enable consumers to access, review, and manage their consolidated health records through a Patient Portal account. Per the FTC, to populate the Patient Health Report, PaymentsMD obtained the sensitive health information of consumers registering for the Patient Portal from health insurance plans, pharmacies, and a medical testing lab, without appropriate authorization from those consumers. The FTC concluded that many consumers registering for the Patient Portal were unaware that respondent would seek to collect their sensitive health data. In fact, the FTC accused PaymentsMD of having presented, directly or indirectly, expressly or by implication, that the authorizations were to be used exclusively to provide the free Patient Portal billing history service for which consumers were registering. PaymentsMD settled with the FTC, among the terms agreed to was that PaymentsMD agreed to destroy the sensitive health information it collected related to the service.[111]

In third example, is the FTC settlement with X-Mode Social and its successor Outlogic.[112] The FTC accused X-Mode Social of failing to ensure that users of its own apps, were fully informed about how their location data would be used.[113] While X-Mode's consumer notices and the consumer notices it provided to third-party app publishers, disclosed certain commercial uses of consumer location data, X-Mode failed to inform consumers that it would be selling data to government contractors for national security purposes. X-Mode also had little or no control over downstream uses of the data it sold. In at least two known instances, X-Mode sold location data to customers who violated contractual restrictions limiting the resale of such data. As part of this settlement, Outlogic as the successor, agreed to delete or destroy all the location data it previously collected, and any products produced from this data unless it obtains consumer consent or ensures the data has been deidentified or rendered non-sensitive. It also agreed to develop a supplier assessment program to ensure that companies that provide location data obtained

[110] https://www.ftc.gov/system/files/documents/cases/141201paymentsmdcmpt.pdf
[111] https://www.ftc.gov/system/files/documents/cases/150206paymentsmddo.pdf
[112] https://www.ftc.gov/system/files/ftc_gov/pdf/X-Mode-D%26O.pdf
[113] https://www.ftc.gov/system/files/ftc_gov/pdf/X-ModeSocialComplaint.pdf

informed consent from consumers for the collection, use and sale of the data or stop using such information.[114]

This flow of data to data brokers and third parties is not only present in the U.S.A. In the EU, for example, the CNIL fined the French data broker Tagada Media €75,000 for collecting prospective customers' data without any legal basis for processing personal data as required by Article 6 of the GDPR. [115] Tagada Media's forms for data collection were found to be misleading and lacked a clear opinion to refuse data collection. This collected data was sent to third parties for commercial prospecting, however the forms used do validly secure consent in compliance with the requirements of the GDPR. The CNIL has also fined other companies for failing to properly secure consent or other legitimate basis for processing personal information, such as Foriou,[116] and Hubside.Store.[117]

5.3 Sale of Personal Data and Data Brokers

The sale and exchange of data is global in scope, yet laws and regulations on the sale of personal data and data brokers vary widely if there are any specific laws at all in each jurisdiction. In the EU, the GDPR is very limited on provisions regarding the selling and trade of personal data. Generally, consent and explicit consent may be used as the legal basis for said sales. However, Ruschemeier (2023) argues that most data brokers are not compliant with the GDPR, due to the fundamental problems of obtaining informed consent in digital environments and the need to balance this against the legitimate interest as a basis for the lawful processing of personal data. In comparison to the United States, there is no federal regulation that comprehensively regulates the sale of personal data. Muralidhar & Palk (2020) previously highlighted the unusual discrepancy between strong privacy obligations for the government and nearly nonexistent privacy obligations for the data brokers in the United States. Nonetheless, regulations have

[114] https://www.ftc.gov/system/files/ftc_gov/pdf/X-Mode-Social-ACCO.pdf
[115] https://www.cnil.fr/en/data-brokers-tagadamedia-fined-eu75000
[116] https://www.cnil.fr/en/commercial-prospecting-foriou-fined-eu310000
[117] https://www.cnil.fr/en/commercial-prospecting-hubsidestore-fined-eu525000

come into force in recent years as California,[118] Nevada,[119] Virginia,[120] Colorado,[121] Connecticut,[122] and Utah[123] have laws that specifically regulate the sale of personal data mostly giving consumers the right to opt out of said sale and even sharing in the case of California.

To illustrate how this regulation of personal data sales works, section 1798.120 of the California Privacy Rights Act (CPRA) which amended the CCPA, expanded California users' right to opt-out expands to the "sharing" as well as the "selling" of their personal information. Section 1798.135 of the CPRA enumerates methods for limiting the sale, sharing, and use of personal information and use of sensitive personal information. This is made stronger by the prohibition on retaliation following opt out rights found in Section 1798.125 of the CPRA. Sharing in the context of the CPRA is broadly defined per Section 1798.12 as ""*Share," "shared," or "sharing" means sharing, renting, releasing, disclosing, disseminating, making available, transferring, or otherwise communicating orally, in writing, or by electronic or other means, a consumer's personal information by the business to a third party for cross-context behavioral advertising, whether or not for monetary or other valuable consideration, including transactions between a business and a third party for cross-context behavioral advertising for the benefit of a business in which no money is exchanged."*

California goes a step further than other states, as the California Delete Act provides that the California Privacy Protection Agency (CPPA) create and maintain an internet website where the information provided by data brokers will be accessible to the public.[124] These brokers will be required to disclose the types of personal information they collect, how they use it, and with whom they share it. The CPPA must establish an accessible deletion mechanism or platform that allows a consumer, through a single verifiable consumer request mechanism to notify every data broker that maintains any personal

[118]https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5
[119] https://www.leg.state.nv.us/Division/Legal/LawLibrary/NRS/NRS-603A.html
[120] https://lis.virginia.gov/cgi-bin/legp604.exe?212+ful+CHAP0036+pdf
[121] https://leg.colorado.gov/sites/default/files/2021a_190_signed.pdf
[122] https://www.cga.ct.gov/2022/act/Pa/pdf/2022PA-00015-R00SB-00006-PA.PDF
[123] https://le.utah.gov/xcode/Title13/Chapter61/13-61.html
[124] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB362

information delete any personal information related to that consumer held by the data broker or associated service provider or contractor, by 01 January 2026. The CPPA also has increased authority to investigate and enforce the Delete Act's provisions.

In the EU, data brokers are not regulated differently from other personal information controllers. In the US, states such as Vermont,[125] Nevada, [126] Texas, [127] Oregon[128] and California[129] have specific regulations on data brokers. Nevada's Revised Statutes Chapter 603A requires that qualifying data brokers must establish a designated request address through which a Nevada consumer can ask to opt out of the sale of their covered information. California, Oregon, Texas, and Vermont require the registration of data brokers. This may change soon as Bureau of Consumer Financial Protection requested the public for Information Regarding Data Brokers and Other Business Practices Involving the Collection and Sale of Consumer Information last March 2023.[130]

6. Regulatory Gaps and Possible Solutions

Oberski & Kreuter (2020) have proposed that de-identification policy should be like cybersecurity policy, since perfect, impregnable security is impossible, it follows that de-identification and data security policy should minimize the risk of breaches and other failures down to acceptably low levels. They have pointed out that sharing even a de-identified dataset involves a non-zero chance of re-identifying or learning something new about people within a de-identified dataset. Ruohonen & Mickelsson (2023) have voiced the concern that state-of-the-art privacy-preserving methods mentioned in the EU DGA cannot prevent de-anonymization and re-identification by efficient algorithms for de-anonymization and re-identification of data subjects. Moreover, the efficiency of these algorithms seems to be increasing with advances in machine learning and artificial intelligence. Oberski & Kreuter (2020) state that the

[125] https://legislature.vermont.gov/statutes/fullchapter/09/062
[126] https://www.leg.state.nv.us/Division/Legal/LawLibrary/NRS/NRS-603A.html#NRS603ASec345
[127] https://statutes.capitol.texas.gov/Docs/BC/htm/BC.509.htm
[128] https://olis.oregonlegislature.gov/liz/2023R1/Downloads/MeasureDocument/HB2052/Enrolled
[129] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB362
[130] https://files.consumerfinance.gov/f/documents/cfpb_request-for-information_data-brokers_2023-3.pdf

regulatory goal should be ensuring the widest number of data custodians reduce such risks to the lowest possible degree, while still retaining the utility of sharing data. Closing the regulatory gaps regarding dark patterns would contribute to this regulatory goal.

The Office of the Victorian Information Commissioner (2022) already pointed out information available to law-abiding researchers today is the absolute minimum that might be available to a determined attacker seeking to re-identify seemingly anonymized data sets, now or in the future. Thus, disclosure risks could be significantly reduced if individuals knew where their personal data was going, who was using it, and what purposes it was being used for. Equally important would be the ability to easily control one's personal data without being the subject of manipulation via dark patterns. This is easier said than done as research conducted by Di Geronimo et. al. (2020) shows that 55% of users did not spot malicious designs in applications containing dark patterns, 20% were unsure, and the remaining 25% found a malicious design.

Dark patterns sharpen the criticism of the current notice and consent based framework for the protection of privacy (Cate & Mayer-Schönberger 2013, Hull 2014 Nehf 2019, and Nouwens et. al, 2020). It has been close to a decade since the NITDRP (2016) stated that it is unfair to place the burden individuals with understanding all the legal, privacy, or ethical implications the of sharing data that being used in new and unanticipated ways in the big data era. Better and systematic regulation of dark patterns would help in reducing disclosure risk for the simple reason that individuals would be able to exercise more control on the spread of their personal information.

Regulatory action has been sparse at best, even cursory searches show the vast amount of personal information for sale, a lot of appears to have been gathered by dark patterns. The OECD (2022) finds that there could be gaps in the law, in available evidence, or in enforcement capacity since enforcement cases predominantly relate to a few dark patterns commonly recognized by regulators. There seems to be a whack-a-mole application of regulations since for every action taken against

companies and brokers for using dark patterns such as PaymentsMD, there are numerous other entities such as Life360,[131] Anomaly Six,[132] and Safeguard,[133] whose similar actions have been reported in media but do not seem to have merited investigation let alone enforcement from regulators.

Companies using dark patterns seem to be incentivized by the market to do so (OECD 2022 & Runge et. al. 2023). Thus, even if there were more comprehensive regulations, push back from companies and data brokers that have thrived in the current big data landscape is expected such as the current litigation questioning the constitutionality of the CAADC. It is one thing for a research team to declare the existence of a dark pattern. It is an altogether different animal for a regulator to determine its existence and implement corrective measures. Understanding that new laws or regulatory frameworks are needed is different from marshaling the power to enact them, let alone the will to adequately enforce them. This is best demonstrated by the light enforcement found by Mahoney (2020) of the Do Not Sell provisions of the CCPA by registered data brokers in California. More troubling was that this study showed some brokers seemingly resorted to dark patterns in an effort to get consumers not to exercise their right not to have their data sold (Mahoney 2020).

Given the non-rival nature of personal information, even the current regulations do not seem to be a viable solution to all the personal data that was already gathered due to dark patterns. Assuming that it can be proven that the personal data was sourced due to a dark pattern, it seems untenable for regulators to going after each and every buyer of this personal data. Many of these buyers relied on the current lack of regulation and would probably be considered to have acted in good faith in arms-length commercial transactions. The dark pattern sourced data in these datasets were allowed to build up over years with scant regulation. It would be wishful thinking that a few laws and occasional regulatory action would instantly remove them from circulation. Nonetheless, data owners need to know where and how

[131] https://theintercept.com/2022/04/22/anomaly-six-phone-tracking-zignal-surveillance-cia-nsa/
[132] https://themarkup.org/privacy/2021/12/06/the-popular-family-safety-app-life360-is-selling-precise-location-data-on-its-tens-of-millions-of-user
[133] https://www.eff.org/deeplinks/2022/05/safegraphs-disingenuous-claims-about-location-data-mask-dangerous-industry

much dark pattern sourced data there is, as there is a risk that it could dirty their otherwise clean de-identified data sets. Knowing how much dark pattern sourced data is out there is especially important for research where current protection methods such as differential privacy are not appropriate such as demographers, redistricting analysts or immigration (Ruggles et. al. 2019, and Groshen &Goroff 2022).

From the perspective of individuals, Hotz et. al. (2022) posit that the individual will care a great deal about small increases in the disclosure risk if the probability of disclosure is already high but may not be bothered by even a large relative increase in risk from data release if the probability of disclosure remains low in absolute terms after release. Laws such as the EU DSA mandate detailed transparency and privacy reporting of VLOPs and VLOSEs with third party audits. The California Delete Act also mandates more transparency from data brokers regarding the sale of personal information. These reports in addition to sweeps conducted should give data owners and individuals a better idea of where their data is going and appropriate steps to take if these data flows are contrary to their best interests. Hopefully this would keep the probability of disclosure low in absolute terms. Ultimately these mandated reports can show policy makers the best possible use of limited regulatory resources, and individuals if they are at a substantial risk of being re-identified.

Going beyond closing regulatory gaps relative to dark patterns, it might be best to frame the balancing the utility of sharing data with researchers and the privacy rights of individuals within the context of the Belmont Principle of beneficence.[134] Beneficence is an obligation: (1) to do no harm and (2) maximize possible benefits and minimize possible harms. Integrating the principle of beneficence to a potential remodeling de-identification law and policy to minimize the risk of breaches and other failures down to acceptably low levels, should give individuals redress in case of privacy breaches due to re-identification in released datasets. It can be argued it is much harder to achieve beneficence due to

[134] https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

individuals lack of awareness of privacy risks due to unregulated or lightly regulated dark patterns in the context of big data.

In this regard, government agencies and research institutions can explore offering reimbursement for privacy losses or help avoiding these losses such as legal assistance in case of identity theft. There can be multiple funding options for this, such as fees from data brokers or even pooled money from fines where dark patterns are prohibited. This could be similar to how subjects in medical research are referred to treatment centers after the conclusion of research participation can help avoid unintended harm.

Questions on how best to implement these rules will vary from country to country. Nonetheless, researchers and policy makers have a clear duty to continue to explore and publicly discuss how to approach these questions in the spirit of transparency and the furtherance of beneficence. Scientific and economic research must not be unduly hindered even if we have yet to find the perfect answers to these questions. We may never find perfect answers rather we can hopefully adopt workable solutions that are most appropriate for their age and context. Technological progress and societal norms are by their very nature dynamic, and thus discussions on balancing these competing needs must be done on a regular basis.

7. Conclusion

This exploratory study highlights the pervasiveness of dark patterns and its effect on disclosure risks. Dark patterns are by no means the only factor in the ongoing debate between protecting privacy and utilizing datasets in the age of big data, but they should still be considered in view of their prevalence and scant regulation. Understanding the gaps in the regulations of these dark patterns should be a key subject for researchers and policy makers as such knowledge will enable society at large better utilize the enormous and varied data sets that fuel big data.

References

Abowd, J. M., & Schmutte, I. M. (2019). An Economic Analysis Of Privacy Protection And Statistical Accuracy As Social Choices. the American Economic Review, 109(1), 171–202. https://doi.org/10.1257/aer.20170627

Abowd, J.M., Adams, T., Ashmead, R., Darais, D., Dey, S., Garfinkel, S.L., Goldschlag, N., Kifer, D., Leclerc, P., Lew, E., Moore, S., Rodr'iguez, R.A., Tadros, R.N., & Vilhuber, L. (2023). The 2010 Census Confidentiality Protections Failed, Here's How and Why. ArXiv, abs/2312.11283.

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2022).Too Much Data: Prices and Inefficiencies in Data Markets. American Economic Journal: Microeconomics, 14 (4): 218-56. DOI: 10.1257/mic.20200200

Acquisti, A., Adjerid, I., & Brandimarte, L. (2013). Gone in 15 seconds: The limits of privacy transparency and control. IEEE Security & Privacy, 11(4), 72–74. https://doi.org/10.1109/msp.2013.86.

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy (March 8, 2016). Journal of Economic Literature, 54(2): 442-492 (2016) http://doi.org/10.1257/jel.54.2.442, Sloan Foundation Economics Research Paper No. 2580411, http://dx.doi.org/10.2139/ssrn.2580411

Athey, S., Catalini C., and Tucker, C. (2017). The digital privacy paradox: Small money, small costs, small talk. Technical report, National Bureau of Economic Research https://www.nber.org/system/files/working_papers/w23488/w23488.pdf

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., and Turner, E. (2019). Americans and Privacy: Concerned, Confused and Feeling Lack of Control over their Personal Information. Pew Research Center. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Bahga, A., & Madisetti, V. K. (2016). Big Data Science & Analytics: A Hands-On Approach. https://dl.acm.org/citation.cfm?id=2948756

Barth-Jones, D. C. (2012). The "Re-Identification" of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2076397

Beraja, M., Kao, A., Yang, D. Y., & Yuchtman, N. (2023). AI-tocracy. *The Quarterly Journal of Economics*, *138*(3), 1349–1402. https://doi.org/10.1093/qje/qjad012

Blake, T., Moshary, S., Sweeney, K., & Tadelis, S. (2021, July). Price Salience and Product Choice. Marketing Science, 40(4), 619–636. https://doi.org/10.1287/mksc.2020.1261

Bösch, C., Erb, B., Kargl, F., Kopp, H., & Pfattheicher, S. (2016). Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*, *2016*(4), 237–254. https://doi.org/10.1515/popets-2016-0038

Cate, F. H. and Mayer-Schönberger, V. (2013) Notice and Consent in a World of Big Data". *Articles by Maurer Faculty*. 2662. https://www.repository.law.indiana.edu/facpub/2662

Cremin, C. J., Dash, S., & Huang, X. (2022). Big data: Historic advances and emerging trends in biomedical research. *Current Research in Biotechnology*, *4*, 138–151. https://doi.org/10.1016/j.crbiot.2022.02.004

Culnane, C., Rubinstein, B. & Teague, V. (2017). Health Data in an Open World. https://doi.org/10.48550/arXiv.1712.05627

Culnane, C., Rubinstein, B.I., & Teague, V. (2019). Stop the Open Data Bus, We Want to Get Off. ArXiv, abs/1908.05004.

Culnane, Chris & Rubinstein, Benjamin & Watts, David. (2020). Not fit for Purpose: A critical analysis of the 'Five Safes'. https://doi.org/10.48550/arXiv.2011.02142

Diebold, F. X. (2012). On the Origin(s) and Development of the Term "Big Data." PIER Working Paper No. 12-037, http://dx.doi.org/10.2139/ssrn.2152421

Di Geronimo, L., Braz, L., Fregnan, E., Palomba, F., & Bacchelli, A. (2020. UI Dark Patterns and Where to Find Them: A Study on Mobile Applications and User Perception. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376600

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017, March 7). Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application*, *4*(1), 61–84. https://doi.org/10.1146/annurev-statistics-060116-054123

European Commission, Directorate-General for Justice and Consumers, Lupiáñez-Villanueva, F., Boluda, A., Bogliacino, F. et al. (2022) *Behavioural study on unfair commercial practices in the digital environment : dark patterns and manipulative personalisation : final report*. Publications Office of the European Union. https://data.europa.eu/doi/10.2838/859030

Farzanehfar, A., Houssiau, F., & De Montjoye, Y. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns*, *2*(3), 100204. https://doi.org/10.1016/j.patter.2021.100204

Forbrukerrådet. (2018a). Deceived by Design. https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf

Forbrukerrådet. (2018b). Every Step you Take. https://storage02.forbrukerradet.no/media/2018/11/27-11-18-every-step-you-take.pdf

Gray, C. M., Santos, C., Bielova, N., & Mildner, T. (2023). An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.09640

Groshen, E. L., & Goroff, D. (2022). Disclosure Avoidance and the 2020 Census: What Do Researchers Need to Know? *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.aed7f34f

Hasan, M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00291-z

Henriksen-Bulmer, J., & Jeary, S. (2016, December). Re-identification attacks—A systematic literature review. *International Journal of Information Management*, *36*(6), 1184–1192. https://doi.org/10.1016/j.ijinfomgt.2016.08.002

Heffetz, O. (2022). What Will It Take to Get to Acceptable Privacy-Accuracy Combinations? . *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.5d9b1a8d

Heffetz, O., & Ligett, K. (2013). Privacy and Data-Based Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2324830

Hidaka, S., Kobuki, S. Watanabe, M., and Seaborn, K. (2023). Linguistic Dead-Ends and Alphabet Soup: Finding Dark Patterns in Japanese Apps. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 3, 1–13. https://doi.org/10.1145/3544548.3580942

Hotz, V. J. & Slanchev, V. (2017). "Designing Consent Protocols to Link Sensitive Health and Administrative Records in Social Science Surveys: Phase I" Accessed at https://public.econ.duke.edu/~vjh3/working_papers/ConsentProject.pdf

Hotz, V. J., Bollinger, C. R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, A., & Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(31). https://doi.org/10.1073/pnas.2104906119

Hotz, V. J., & Salvo, J. (2022). A Chronicle of the Application of Differential Privacy to the 2020 Census. *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.ff891fe5

Hull, G. (2015). Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. *Ethics and Information Technology*, *17*(2), 89–101. https://doi.org/10.1007/s10676-015-9363-z

Ichihashi, S. (2020). Non-competing data intermediaries. Staff Working Papers 20-28, Bank of Canada. https://doi.org/10.34989/swp-2020-28

Johnson, G. (2022). Economic Research on privacy regulation: Lessons from the GDPR and beyond. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4290849

Jones, C. I., & Tonetti, C. (2020, September 1). Nonrivalry and the Economics of Data. *American Economic Review*, *110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330

Jouhki, J., Lauk, E., Penttinen, M., Sormanen, N., & Uskali, T. (2016, October 10). Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication*, *4*(4), 75–85. https://doi.org/10.17645/mac.v4i4.579

Kallioniemi, P. (2022). Facebook's dark pattern design, public relations and internal work culture. *DOAJ (DOAJ: Directory of Open Access Journals)*. https://doi.org/10.34624/jdmi.v5i12.28378

Keller, S. A., & Abowd, J. M. (2023, March 15). Database reconstruction does compromise confidentiality. *Proceedings of the National Academy of Sciences*, *120*(12). https://doi.org/10.1073/pnas.2300976120

Kelly, D., & Burkell, J. (n.d.). *Documenting Privacy Dark Patterns: How social networking sites influence users' privacy choices*. Scholarship@Western. https://ir.lib.uwo.ca/fimspub/376

Kern, J., Fabian, B., & Ermakova, T. (2018). Experimental Privacy Studies - An Initial Review of the Literature. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3171929

Kitkowska, A. (2023). The Hows and Whys of Dark Patterns: Categorizations and Privacy. In: Gerber, N., Stöver, A., Marky, K. (eds) Human Factors in Privacy Research. Springer, Cham. https://doi.org/10.1007/978-3-031-28643-8_9

Kollmer, T. & Eckhardt, A. (2022). Dark Patterns: Conceptualization and Future Research Directions. Business & Information Systems Engineering. 65. 10.1007/s12599-022-00783-7.

Kowalczyk, M., Gunawan, J. T., Choffnes, D., Dubois, D. J., Hartzog, W., & Wilson, C. (2023, April 19). Understanding Dark Patterns in Home IoT Devices. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581432

Leiser, D. M. (2020). "Dark Patterns": The Case for Regulatory Pluralism. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3625637

Liu, Z., Sockin, M., & Xiong, W. (2023). Data Privacy and Algorithmic Inequality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4448301

Lohr, S. (2012, February 11) The Age of Big Data, *New York Times*, https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Luguri, J. B., & Strahilevitz, L. (2021). Shining a light on dark patterns. *the Journal of Legal Analysis*, *13*(1), 43–109. https://doi.org/10.1093/jla/laaa006

Mahoney, M., (2020) California Consumer Privacy Act: Are Consumers' Digital Rights Protected? Consumer Reports Digital Lab, http://advocacy.consumerreports.org/wp-content/uploads/2020/09/CR_CCPA-Are-Consumers-Digital-Rights-Protected_092020_vf.pdf

Marti, D., Lin, F. Scharwtz, M. and Fahs, J. (2024). Who shares your information with Facebook: Sampling the Surveillance Economy in 2023. https://innovation.consumerreports.org/wp-content/uploads/2024/01/CR_Who-Shares-Your-Information-With-Facebook.pdf

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale. *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 1–32. https://doi.org/10.1145/3359183

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What makes a dark pattern. . . dark? *arXiv (Cornell University)*. http://arxiv.org/abs/2101.04843

Mazzei, M. & Noble, D. (2019). Big Data and Strategy: Theoretical Foundations and New Opportunities. 10.5772/intechopen.84819.

McGeeney, K., Kriz, B., Mullenax, S., Kail, L., Walejko, G., Vines, M., Bates, N., & García Trejo, Y. (2019). *2020 Census Barriers, Attitudes, and Motivators Study survey report: A new design for the 21st century*. U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/program-management/finalanalysis-reports/2020-report-cbams-study-survey.pdf

Mhaidli, A.H., Zou, Y., & Schaub, F. (2019). *"We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks.* In Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security (SOUPS'19). USENIX Association, USA, 225–244. https://dl.acm.org/doi/10.5555/3361476.3361493

Muralidhar, K., & Palk, L. (2018). A free ride: Data Brokers'Rent-Seeking Behavior and the Future of data Inequality. *Vanderbilt Journal of Entertainment & Technology Law*, *20*(3), 779. https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1097&context=jetlaw

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings - IEEE Symposium on Security and Privacy/Proceedings of the . . . IEEE Symposium on Security and Privacy*. https://doi.org/10.1109/sp.2008.33

Narayanan, A., Mathur, A., Chetty, M., & Kshirsagar, M. (2020, April 30). Dark Patterns: Past, Present, and Future. *Queue*, *18*(2), 67–92. https://doi.org/10.1145/3400899.3400901

Nehf, J.P. (2019). The Failure of 'Notice and Consent' as Effective Consumer Policy. LSN: Regulation of Contracting Private Parties (Topic). http://dx.doi.org/10.2139/ssrn.3440816

Nissenbaum, H. (2009). *Privacy in context, technology, policy, and the integrity of social life*. Stanford University Press.

Nouwens, M., Liccardi, I., Veale, M., Karger, D., & Kagal, L. (2020, April 21). Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376321

Nycyk, M. (2021). Ethical Use of Informant Internet Data: Scholarly Concerns and Conflicts. *Journal of Digital Social Research*, *4*(1), 1-22. https://doi.org/10.33621/jdsr.v4i1.88

Oberski, D. L., & Kreuter, F. (2020). Differential Privacy and Social Science: An Urgent Puzzle. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.63a22079

Ochoa, S.J., Rasmussen, J.C., Robson, C., & Salib, M. (2002). Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study.

OECD (2022), "Dark commercial patterns", *OECD Digital Economy Papers*, No. 336, OECD Publishing, Paris, https://doi.org/10.1787/44f5e846-en.

The Office of the Victorian Information Commissioner (2022). The Limitations of De-Identification – Protecting Unit-Record Level Personal Information. https://ovic.vic.gov.au/privacy/resources-for-organisations/the-limitations-of-de-identification-protecting-unit-record-level-personal-information/

Ohm, P. (2009). Broken Promises of privacy: Responding to the surprising failure of anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010, U of Colorado Law Legal Studies Research Paper No. 9-12, Available at SSRN: https://ssrn.com/abstract=1450006

Owens, K., Gunawan, J., Choffnes, D., Emami-Naeini, P., Kohno, T., & Roesner, F. (2022, September 29). Exploring Deceptive Design Patterns in Voice Interfaces. *Proceedings of the 2022 European Symposium on Usable Security*. https://doi.org/10.1145/3549015.3554213

Pereira, T. F., Aranha, V. J., Waldvogel, B. C., Da Costa, A. M., & Fregnani, J. H. T. G. (2023). Deterministic linkage for improving follow-up time in a Brazilian population-based cancer registry. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-31303-6

Press, G. (2019, July 17). A very short history of big data. Forbes. https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/?sh=116b2a9865a1

Reilly, B. C. (2015). Doing More with More: The Efficacy of Big Data in the Intelligence Community. *American Intelligence Journal*, *32*(1), 18–24. http://www.jstor.org/stable/26202099

Rocha, A., Azevedo, L. F., Silva-Cardoso, J., Allison, T. G., & Freitas, A. (2019). Internal Deterministic Record Linkage Using Indirect Identifiers For Matching Of Same-Patient Hospital Transfers And Early Readmissions After Acute Coronary Syndrome In A Nationwide Hospital Discharge Database: A Retrospective Observational Validation Study. *BMJ Open*, *9*(12), e033486. https://doi.org/10.1136/bmjopen-2019-033486

Rocher, L., Hendrickx, J. M., & De Montjoye, Y. (2019). Estimating The Success Of Re-Identifications In Incomplete Datasets Using Generative Models. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10933-3

Ruohonen, J. & Mickelsson, S. (2023). Reflections on the Data Governance Act. Digital Society. 2. 10.1007/s44206-023-00041-7.

Rubinstein, I. (2013). Big data: the end of privacy or a new beginning? *International Data Privacy Law*, *3*(2), 74–87. https://doi.org/10.1093/idpl/ips036

Ruggles, S., (2018). "Implications of Differential Privacy for Census Bureau Data and Scientific Research." Minnesota Population Center Working Paper 2018-6 https://assets.ipums.org/_files/mpc/wp2018-06.pdf#%5B%7B%22num%22%3A370%2C%22gen%22%3A0%7D%2C%7B%22name%22%3A%22FitH%22%7D%2C792%5D

Ruggles, S., Fitch, C. A., Magnuson, D. L., & Schroeder, J. P. (2019). Differential Privacy and census Data: Implications for social and economic research. *AEA Papers and Proceedings*, *109*, 403–408. https://doi.org/10.1257/pandp.20191107

Runge, J., Wentzel, D., Huh, J.Y. & Chaney, A. *(2023).* "Dark patterns" in online services: a motivating study and agenda for future research. *Mark Lett* 34, 155–160. https://doi.org/10.1007/s11002-022-09629-4

Ruschmeier, H. (2023). Data brokers and European digital legislation. *European Data Protection Law Review*, *9*(1), 27–38. https://doi.org/10.21552/edpl/2023/1/7

Sala, E., Knies, G., & Burton, J. (2014). Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*, *17*(5), 455–473. https://doi.org/10.1080/13645579.2014.899101

Schmutte, I., and Vilhuber, L. (2020). "Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), *Handbook on Using Administrative Data for Research and Evidence-based Policy*. Accessed at https://admindatahandbook.mit.edu/book/v1.0/discavoid.html on 2024-04-05.

Sherman, J., (2021) Data Brokers and Sensitive Data on U.S. Individuals, available at https://techpolicy.sanford.duke.edu/wp-content/uploads/sites/4/2021/08/Data-Brokers-and-Sensitive-Data-on-US-Individuals-Sherman-2021.pdf

Soe, T. H., Nordberg, O. E., Guribye, F., & Slavkovik, M. (2020, October 25). Circumvention by design - dark patterns in cookie consent for online news outlets. *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. https://doi.org/10.1145/3419249.3420132

Solove, D. J. (2004). *The Digital Person: Technology and Privacy In The Information Age*. New York: New York University Press.; available at https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2501&context=faculty_publications;

Spiekermann, S., Acquisti, A., Böhme, R., & Hui, K. L. (2015, April 29). The challenges of personal data markets and privacy. *Electronic Markets*, *25*(2), 161–167. https://doi.org/10.1007/s12525-015-0191-0

Stoughton, J. W., Whelan, T. J., & Thompson, L. F. (2015). Perceptions of confidentiality in survey research: Development of a scale. ResearchGate. https://www.researchgate.net/publication/269104204_Perceptions_of_confidentiality_in_survey_research_Development_of_a_scale

Susser, D., Roessler, B., & Nissenbaum, H. F. (2018). Online Manipulation: Hidden Influences in a Digital World. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3306006

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1410

Sweeney, L. (2011). Patient identifiability in pharmaceutical marketing data. *Harvard University, Cambridge, MA, WP--1015*. https://dataprivacylab.org/projects/identifiability/pharma1.pdf

Sweeney, L. *(2015). Only You, Your Doctor, and Many Others May Know. Technology Science*. 2015092903. https://techscience.org/a/2015092903/

Sweeney, L. (2018). Patient Identifiability in Pharmaceutical Marketing Data. Carnegie Mellon University. Journal contribution. https://doi.org/10.1184/R1/6625193.v1

Tahaei, M. and Vaniea, K. (2021). “Developers Are Responsible”:
What Ad Networks Tell Developers About Privacy. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 253, 1–11. https://doi.org/10.1145/3411763.3451805

Tahaei, M., Ramokapane, K. M., Li, T., Hong, J. I., & Rashid, A. (2022). Charting App Developers’ Journey Through Privacy Regulation Features in Ad Networks. *Proceedings on Privacy Enhancing Technologies*, *2022*(3), 33–56. https://doi.org/10.56553/popets-2022-0061

Tan, X., Qin, L., Kim, Y. and Hsu, J. (2012) Impact of Privacy Concern in Social Networking Web Sites. Internet Research, 22, 211-233. https://doi.org/10.1108/10662241211214575

Torra, V., & Navarro-Arribas, G. (2023). Attribute disclosure risk for k-anonymity: the case of numerical data. *International Journal of Information Security*, *22*(6), 2015–2024. https://doi.org/10.1007/s10207023-00730-x

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). (Un)informed Consent. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 973–990. https://doi.org/10.1145/3319535.3354212

Utz, C., Amft, S., Degeling, M., Holz, T., Fahl, S., & Schaub, F. (2023). Privacy rarely considered: Exploring considerations in the adoption of Third-Party services by websites. *Proceedings on Privacy Enhancing Technologies*, *2023*(1), 5–28. https://doi.org/10.56553/popets-2023-0002

Wolf LE. Risks and Legal Protections in the World of Big-Data. Asia Pac J Health Law Ethics. 2018 Mar;11(2):1-15. Epub 2018 Mar 31. PMID: 31745539; PMCID: PMC6863510.

Yoo J., Thaler A., Sweeney L., and Zang J. (2018) Risks to Patient Privacy: A Re-identification of Patients in Maine and Vermont Statewide Hospital Data. *Technology Science*. https://techscience.org/a/2018100901/

Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5