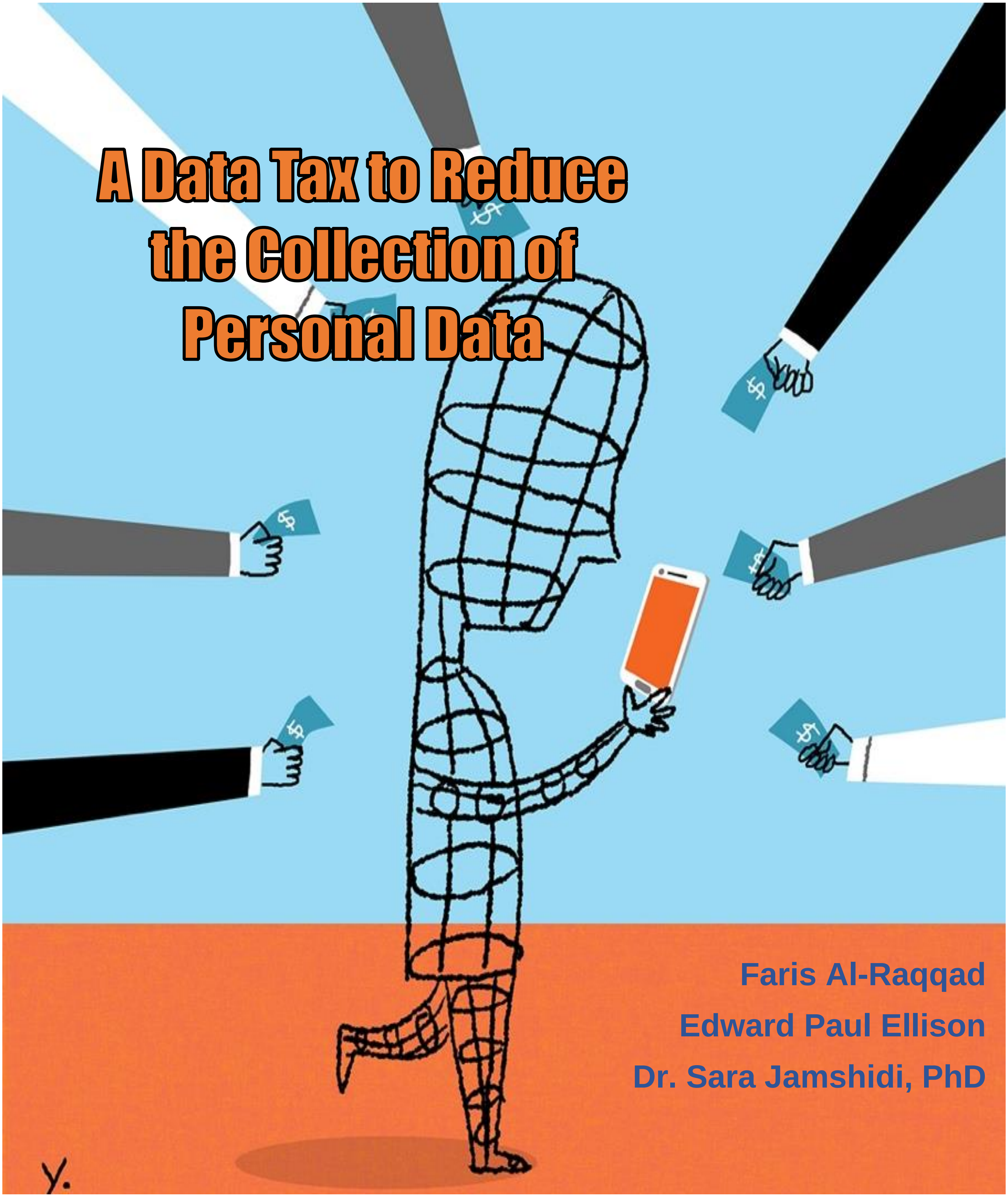
A Data Tax to Reduce the Collection of Personal Data



Faris Al-Raqqad

Edward Paul Ellison

Dr. Sara Jamshidi, PhD

Abstract

The CCPA, CCPR, and GDPR are well intentioned documents that aim to protect data privacy but fall short of their goals because they are fundamentally hard to enforce. A tax is a common method to reduce activity in a market, such as carbon taxes. It is difficult to track the revenue generated from collected data, however, making taxing the market a challenge. In this paper, we instead propose a graded tax on all data transfers. The intention of this is to disincentivize mass data collection by tech companies, like data brokerage firms and social media sites, by taxing the source of their revenue. Data transfers up to (about) 5,000-10,000 gigabytes/month will not be affected by this tax, however, so that the average internet user is not affected. Companies that rely on consumers downloading data, such as streaming services like Netflix, should also be unaffected by their normal operation. The result is that companies will need to be more judicious when collecting data, leading to less data being collected overall. This will lead to increased costs for purposes that serve the public interest such as research. Non-profit entities such as universities should still qualify for exemption from this tax to lessen its impact on society.

Executive Summary

Nowadays, the fact that consumers' personal data is sold by social media companies is accepted as a compromise for being offered these services for free. The scope and sheer volume of personal data being collected, however, is far greater than most would imagine. These ongoing issues of informed consent have been addressed before, either by the companies themselves or some form of legislation. Rapid development of new technologies and algorithms will both greatly increase the amount of information extracted from users' personal data and make it harder for governments to regulate the new practices before damage is done.

There have been attempts to legislate against the intense data collection, such as the GDPR in the EU and the CCPA/CCPR in California. These pieces of legislation rightly target the lack of informed consent on the consumer's part that has been taken for granted and give the consumer the right to opt-out of personal data collection. Despite their intentions, these regulations have often been hard to enforce for multiple reasons.

One major reason these regulations struggle is because the users of internet services are conditioned to accept the Terms of Services and Privacy Policies without fully understanding them. This allows companies like Meta, TikTok, and Instagram to collect data far beyond what the consumers could possibly be aware of, with the consumer consent.

Another issue that legislators run into when trying to limit harmful personal data collection through these regulations is due to these services being international products. If a Californian company accessed a Massachusetts' personal data in a harmful way, and stored it on a server based in Seattle, who has jurisdiction? This issue has historically been more pronounced in the enforcement of the GDPR but is a strong argument for federal adoption of the tax policy discussed in this paper.

Currently, a rapidly developing field of technology is the field of generative AI research; specifically, Microsoft's OpenAI's GPT-4 has been a massive improvement on its predecessor, GPT-3. It requires a massive amount of data, and we are not totally sure about what it was trained on as OpenAI has not been transparent about what is included in its training dataset. Technology that can fabricate convincing video of people, deepfakes, has existed for years and continues to improve thanks to an increase in high-definition, mass collected data. In recent years, "voice cloning" technology has advanced greatly. Scammers only need a few seconds of a loved one's voice to clone it and use it to manipulate some victim into giving them money, or worse, extremely personal information (Hernandez, 2023).

Though it seems like the data economy's meteoric rise in the digital age is inevitable and unstoppable, there is a solution that may seriously slow down the high amount of data collected: a data collection tax. This would be a graded tax, designed to avoid affecting households – a surprisingly simple calculation considering the chasm between consumer download rates and those of companies like Meta and Alphabet. By taxing data collection, companies will have to be more cautious about how they collect data or pay for the excise.

Team Biographies

Faris Ahmad Al-Raqqad

Team member

Faris is a senior at Lake Forest College majoring in Computer Science. His interests are in artificial intelligence, database management, and digital ethics. He has worked as a mathematics tutor for K-12 students, held the positions of Treasurer, Vice President, and President of his fraternity ΛXA, and is interested in pursuing a career in database management. He aims to assist policy makers in stopping harmful personal data collection.

Edward Paul Ellison

Former team member

My name is Paul Ellison. I live in Winnetka, and am a junior at Lake Forest College, a liberal arts university. My interests are history (with special emphasis on medieval history), physics (particularly concerning the quantum level), and religion. I am majoring in history, as it is a topic I am very familiar with.

My aim for a public policy is one which protects consumers from having their data collected without consent. This practice is problematic because it increases violations of privacy, and we aren't free without privacy. Until the United States prioritizes the privacy of its people, they cannot claim to be the land of the free.

Sara Jamshidi

Team advisor

Sara is an assistant professor of Computer Science and Mathematics and director of the Applied Data Center at Lake Forest College. Her training is in computational algebra, algebraic statistics, category theory and nonstandard probability. As a graduate student, she also worked in AI and collaborated with the US Air Force and US Navy. In 2017, she won the Sydney Drell Academic Award from the Intelligence and National Security Alliance. Her current research focuses on machine learning, with special focus on using ML techniques to further work applied computational algebra.

Introduction

Jaron Zepel Lanier, computer scientist and philosopher, has famously criticized "free" elements of the internet such as social media (Lanier 2018). His primary argument is that when a service like Facebook is "free," it is because users are the product being sold. Platforms such as Facebook, Google, Tinder, LinkedIn, and others primarily generate revenue through *targeted* ads. The implementation of targeted ads involves collecting personal data of a consumer and algorithmically extrapolating from that information how best to advertise to the said consumer.

The sheer volume of data and the extent to what is directly known or extrapolated from this data would surprise many. French author, Judith Duportail, exercising her rights under French law, requested a copy of the data collected about her from her use of Tinder. After months of communication, she successfully obtained over 700 pages of personal data collected about her. The report listed many personal pieces of information, including predictions of what times of day she was most lonely (Duportail 2019).

Although many individuals know their data is collected, most are unaware of what information is collected and the extent. The chart below details respondent data collected by Morey et al. in 2015.

Information	*Social Networks Friend's List*	*Location*	*Web searches*	*Communication history, like chat logs*	*IP addresses*	*Web-surfing history*
Percent Aware	**27%**	**25%**	**23%**	**18%**	**17%**	**14%**

All this amounts to a serious issue of informed consent with internet users. Although most applications and account-based sites involve users agree to terms and services, it is clear these legal documents do little to educate users. More importantly, if users are empowered by legislation—like those discussed in the next section—to be able to obtain the personal data collected from them, the process is tedious and the report provided is much too large for the average person to comb though, as shown by Duportail's work.

The GDPR and CCPA

The General Data Protection Regulation, or GDPR, is a regulation that attempts to protect the privacy of individuals who use any online services or applications that operate in the EU (Crutzen 2019). It has had many positive impacts on how terms and services documents are written for most applications and account-based sites. For example, paragraph 3 of Section 13 requires that automated messages be clear in who the sender is and provide a way to opt-out (*GDPR 2016*). It also creates several distinctions between the different applications of data to better target bad actors. Sections 15-22 go on to provide many common-sense guidelines for what should be included in an acceptable Terms of Services Agreement: such as a clear explanation of how data is used, and a user-friendly way to opt-out and/or have their personal data deleted. Additionally, in the event of a data breach, the holder of the compromised user data is expected to notify the affected parties "in a timely manner." Different regulations apply to different uses, such as in the cases of research, business, and automated business. Several supervisory institutions, like the Data Protection Commission in Ireland, were founded to find infractions of the GDPR and litigate the responsible parties.

The GDPR provides a legislative framework for digital service providers to abide by, complete with governmental bodies that are tasked with making sure this framework is upheld. It has also gone on to inspire similar legislation in America's tech capital: the California Consumer Privacy Act, or CCPA, and the later CPRA, or the California Privacy Rights Act, which were enacted on the first day of 2020 and 2023 respectively.

Like the GDPR, the CCPA was written into law to create protections for Californian consumers' personal data, which is data that can be associated with or connected to them (e.g., location data). It gave Californian consumers the right to know who has their personal data, what data they have, and the right to opt-out or object to the data collection wholesale. If a data breach were to occur, the holder of the compromised data would be expected to notify the affected parties within two weeks of becoming aware of the breach. The CPRA granted two additional rights to consumers: the right to correct inaccurate information, and the right to limit how their sensitive personal information (e.g., financial information, demographic data, biometric data) is stored and shared. These two acts are enforced by the California Attorney General, but it invests the California Privacy Protection Agency with "full administrative power, authority, and jurisdiction to implement and enforce" them (California, 2020). Enforcement of the CPRA will begin on July 1st, 2023.

While these regulations are well intentioned, there is little evidence that they are as effective as they could be. The DPC's latest annual report states that they prosecuted only two companies successfully for sending unsolicited marketing material. 38 complaints were filed under the Law Enforcement Directive, and 58 were 'concluded.' These high-profile fines include €492 million fined to Facebook for failing to be as transparent as was expected, and €210 million fined to Instagram for failing to protect children's data. However, an ongoing concern within the DPC is receiving these fines. Because of the interplay between different nations that becomes necessary in enforcing the GDPR, it seems to be extremely difficult to finalize a decision, impose fines, and enable them to be transferred. In its report, the DPC states that the GDPR has "created something of a legal maze that requires constant navigation, building an ever more complex landscape for litigators," (Hughes, 2023).

This is in large part due to the international nature of online business, which both makes it difficult to narrow down where infractions occurred and unclear who the litigator ought to be. An equally problematic issue is the mechanism by which the DPC is made aware of infractions: consumer complaints. As we will outline in the next section, while this makes logistical sense, consumers simply cannot be expected to understand and be invested enough in the GDPR's rulings and how that relates to each ToS/Privacy Policy they may interact with. The information is too technical, specific, and vast for awareness of it to be reasonably expected of the average person.

An even more detrimental issue that these laws and regulations grind against is human nature itself; a study conducted by Eric P. Robinson, an assistant professor at USC's School of Journalism and Mass Communication, found no evidence that the updates to Google and Instagram's Terms of Services documents improved consumer/user understanding of the legalese and rights presented within them (Robinson, 2019). Robinson and his collaborators found that, while reading ToS improved understanding and knowledge of the user's legal rights, the 'simplified' language in these rewritten documents resulted in worse understanding of the rights presented. One possible reason for this is recency bias, meaning that people remember interesting concepts more clearly if they were presented to them recently. This would mean that a more concise and "unreadable" ToS, while confusing, is more likely to stick in a layman's head. Furthermore, most people do not read the Terms of Service nor Privacy Policy; even if a service tries to force someone to read them, they'll often scroll or skim through it until they reach the confirmation at the end. According to Robinson's study, Terms of Services documents are often communicated and interpreted socially, whether it be through observing what others do, or discussions with others about what they think is and is not allowed on a given platform.

Ultimately, legislation like the GDPR and CCPA/CCPR are an important part of any major shift in digital privacy laws and will continue to be essential in the fight for the digital consumer's right to privacy. However, by reflecting on the past seven years, it should become clear that these laws on their own are not enough to govern tech giants, data brokerage firms, and bad actors. They cost a lot of taxpayer money to uphold, and often generate complex legal battles that take years to resolve – and that's only if the data holder who's mishandling personal data gets reported and investigated. Rewrites to the Terms of Service and similar documents, while seemingly helpful, also show no evidence of having a significant impact on the way users interact with and understand their rights regarding digital platforms. Whether it be due to human nature or the ever-evolving technological landscape, regulations alone are not enough to enact real change in how these companies conduct business.

A Data Tax

A mechanism used by governments across the world to limit unfavorable activities in a market is taxation. Historically, these have taken many forms, whether they be sugar taxes which target the amount of soda consumers drink, carbon taxes which aim to reduce emissions caused by fossil-fuels and other polluting energy sources, or cigarette taxes that target tobacco use. In his 2021 paper, *Taxing Data*, Omri Marian, a professor of Law at UC Irvine, the blueprint for a graded data tax, or a "data flow tax,'" is outlined (Marian, 2021). This would be a tax on downstream broadband – or downloads – imposed on all Americans and American corporations. This data tax's tax base, or the metric by which tax is determined, would simply be the amount of data downloaded in gigabytes. In *Taxing Data*, Marian outlines a few clear benefits of a tax system like this, not the least of which is its adaptability.

As mentioned in the previous section, a common pain point for legislators who want to reign in 'big tech' is the internet's ever evolving landscape; by the time a practice is regulated, two new ones have taken its place. This is evident in the data scraping issues that seem to constantly be coming out about Facebook, Google, TikTok, and other digital or online services (Jeffrey, 2023). However, by simply taxing the raw amount of data that the company collects as a commodity, any business that trades personal data for profit will have to conduct cost benefit analyses for what kind of data they collect, how frequently, etcetera.

Another hurdle to creating a "fair" tax is that of multinational enterprises. Many companies offshore different aspects of their business for favorable overall tax rates. For example, in 2019 Apple paid an estimated effective tax rate of 18.1%, Facebook paid 13.1%, and Alphabet, Google's parent company, paid only 8.8% (Future Agenda, 2020).

A data tax may assist fiscal policymakers in combating this, as taxing downloads while they happen will reduce, if not eradicate, concerns about jurisdiction. Cell carriers can track the flow of data, so there is no reason to believe that internet providers and the like could not as well. Moreover, taxing the volume of downloads has the benefit of self-adjustment. He argues,

> *One of the difficulties of the current tax system… …is the mere fact that development happens much more quickly than tax legislation happens… …Instead of trying to adjust the tax each time a new technology appears, taxing the raw commodity that enables technological advances will automatically adjust the tax collected: the more of it is used, the higher the tax (Marian, 561).*

This strength also allows the tax to be created in such a way that it does not affect the average consumer. Since the amount of data collected by these data giants far exceeds that of a standard broadband user, a floor should be introduced to ensure that consumers do not have the cost forwarded onto them. We will explore methods to differentiate which data users should be subject to this tax in the Analysis section.

Analysis

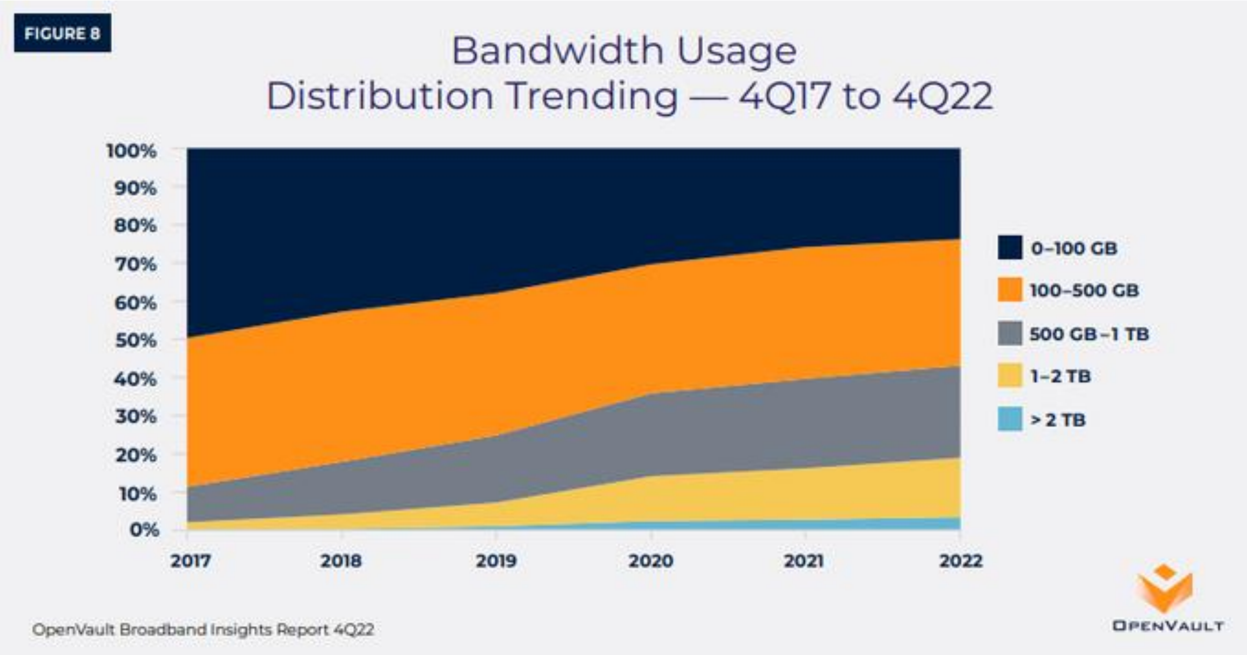
A data collection tax would solve two problems that have plagued fiscal policymakers: multinational conglomerates offshoring revenue to juke taxes, and the unclear economic value of consumer personal data. As explained earlier, it is often difficult to tax or even levy fines against companies like Meta and Instagram for violating consumer privacy due to the internet being an international resource. This has caused a litany of issues, including questions about who should be prosecuting the company, and which state or states get to determine fines.

The DPC's struggle to close some of their largest fines this past year is a clear illustration of this, leaving nearly one billion euros in limbo and draining resources from the taxpayer. This is not an indictment of the GDPR, CCPA/CCPR, nor any similar legislation – rather, it is a sign that the way we think about taxing this industry needs to be amended so it may better reflect how profit is generated. Without a constant incentive to not collect as much data as possible, companies that depend on personal data collection will continue to find ways around regulations, which will always fall behind technological advancement. That is, a data tax, as the earlier Marian quote explains, disincentivizes companies from collecting more data than they are willing, or able, to pay for.

While the exact figures on what this tax should be are beyond the scope of this proposal, there is precedent for what it should look like. Taxes intended to limit harmful commercial activity are sometimes called Pigouvian taxes, named after the economist Arthur Pigou. Pigouvian taxes are often instated not with revenue generation in mind, but with the intent to reduce some activity that is deemed to be against the public interest. If a dollar cost can be attributed to the societal damage caused by some measurable metric, be it tobacco, carbon, or user data, a well-structured tax on it would cause that metric's profiteers to pay it forward.

These taxes can be fairly hit-or-miss, and many factors can affect whether they will be successful in the long run. An example of a successful Pigouvian tax would be cigarette taxes, which according to the American Lung Association, are positively correlated with decreased cigarette use in both adults and minors (ALA, 2022). These taxes target a market that causes uniform harm, cigarettes, and have the explicit goal of reducing the amount of people who smoke; thus, even if the cost is returned to consumers, the net result is a decrease in the amount of people who smoke. Similar reasoning can explain why sugar and carbon taxes have been effective in reducing soda consumption and carbon emissions respectively. The money collected through these taxes can then be used to fund the Federal Trade Commission, better empowering it to regulate tech giants.

A natural concern around this type of tax is how it may affect the average internet user. The simplest solution to this problem is to set a relatively high bar for the tax to even be considered. According to OpenVault's 2022 Q4 report, the average American household's monthly data flow is below 600 gigabytes of data (OpenVault, 2023). This number has been steadily rising for a variety of obvious factors: the increase in people working from home (i.e., Zoom and Microsoft Teams), the prevalence of high-definition video streaming, and the general upwards trend in online activity. Some so-called 'power users' go above this 600GB monthly rate; 18.7% of users use 1000-1999GB, or 1.0-1.9 terabytes, and merely 3.4% use 2TB or more. Most of this data flow is in the form of downloads.



[1]

By comparison, the amount of data flowing through some tech companies dwarfs individual broadband usage by orders of magnitude. While updated and precise figures are difficult to find, the available statistics on the volume of data collected by conglomerates like Meta and Alphabet are astounding. In a 2014 post to Meta's research blog, "Facebook's Top Open Data Problems," Janet Wiener claims that as much as 4 petabytes of data are collected every single day, and the cloud-based storage solution used within the company, the "Nest," can store as much as 400PB (Wiener, 2014). For reference, a petabyte is equivalent to a thousand terabytes, or one million gigabytes. It is difficult to fully comprehend the difference in magnitude, but to better illustrate

[1] Image sourced from OpenVault's "Broadband Insights Report (OVBI)" for Q4 2022.

this point: if each pixel in the image above were worth 600GB, a generous estimation of the data volume downloaded by a typical American household, the image above would only represent a third of the personal user data Meta has at its disposal.

With this in mind, we can safely set the minimum download volume around 5,000-10,000GB, and be confident that consumers won't catch up for a while. It is important to note that this would require frequent surveillance on the government's part; specifically, the tax would need to be tuned to adjust for changes in business practices, the amount of personal data collected, and the amount of data that flows to and from the average consumer. Moreover, if this limit affects the works of nonprofit and governmental agencies, a special clause can be added for institutions with tax-exempt status.

Conclusion

Undergirding the problem of consumer protection in the face of data collection is a much larger philosophical problem facing policy makers. Algorithmic approaches applied to data available on the internet give rise to effects on scales not fully comprehensible to humans. The average internet user does not have a clear sense of what data is collected on them, what types of data is being collected, and the sheer volume of it. More importantly, policy makers are not likely to be fully aware. In order to allow for human oversight of algorithmic processes, we need policies that first and foremost, slow down these processes to scales more manageable by a human.

Ultimately, a data tax has the potential to seriously limit the amount of data collected by multinational entities, as it makes mass data collection vastly more costly to the business. With appropriate pricing, businesses will need to audit data collection methods and perform internal cost-benefit analyses. If smaller, randomized samples of targeted data collection give insights into groups of consumers, business may opt for these types of strategies over blanket data collection as such a strategy may be entirely free. Moreover, opt-out policies like those outlined in the GDPR will be seen as potential cost-saving measures, meaning companies may offer less resistance to such legislation.

Ironically, the antidote to blanket data collection may just be a blanket tax on data transfer.

Bibliography

American Lung Association. “Cigarette & Tobacco Taxes.” *American Lung Association*, 17 Nov. 2022, https://www.lung.org/policy-advocacy/tobacco/tobacco-taxes#:~:text=Every%2010%20percent%20increase%20in,about%20seven%20percent%20among%20youth.

“Consolidated Text: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance).” *Official Journal of the European Union*, 4 Apr. 2016, https://doi.org/02016R0679-20160504.

Crutzen, Rik, et al. “Why and How We Should Care about the General Data Protection Regulation.” *Psychology & Health*, vol. 34, no. 11, 2019, pp. 1347–1357., https://doi.org/10.1080/08870446.2019.1606222.

“Data Taxation.” *Future Agenda*, 21 Jan. 2020, https://www.futureagenda.org/foresights/datataxation/.

Duportail, Judith. (2019) *L'Amour sous algorithme (Love under algorithm)*.

The European Union. (2016) General Data Protection Regulation (GDPR). Retrieved April 9, 2023, from https://gdpr-info.eu/. *Official Legal Text*.

Ghaffary, Shirin. “The Makers of Chatgpt Just Released a New AI That Can Build Websites, among Other Things.” *Vox*, Vox, 15 Mar. 2023, https://www.vox.com/2023/3/15/23640640/gpt-4-chatgpt-openai-generative-ai.

Hernandez, Joe. “That Panicky Call from a Relative? It Could Be a Thief Using a Voice Clone, FTC Warns.” *NPR*, NPR, 22 Mar. 2023, https://www.npr.org/2023/03/22/1165448073/voice-clones-ai-scams-ftc.

Hughes, Ruth, et al. “Highlights of the Data Protection Commission's 2022 Annual Report.” *Lexology*, McCann FitzGerald LLP, 9 Mar. 2023, https://www.lexology.com/library/detail.aspx?g=d8a29ddb-6d61-4e69-8b90-33adda104625.

Jeffrey, Cal. “TikTok and Others Scrape Your Data, Whether You Use Their Apps or Not.” *TechSpot*, TechSpot, 3 Oct. 2022, https://www.techspot.com/news/96187-tiktok-others-scrape-data-whether-you-use-their.html.

Lanier, Jaron. *Ten Arguments for Deleting Your Social Media Accounts Right Now*. Vintage, 2019.

Morey, Timothy; Fortbath, Theodore; Schoop, Allison. (2015) Customer Data: Designing for Transparency and Trust. *Harvard Business Review*.

Robinson, Eric P., and Yicheng Zhu. ‘Beyond “I Agree”: Users’ Understanding of Web Site Terms of Service’. *Social Media + Society*, vol. 6, no. 1, SAGE Publications, Jan. 2020, https://doi.org/10.1177/2056305119897321.

UC Irvine School of Law Research Paper, ser. 2021-17, 26 Feb. 2021. *2021-17*, https://ssrn.com/abstract=3793892.

United States, Congress, State of California. *Cal. Code Regs. Tit. 11, § 7000 - Title and Scope*, 2020.

Wiener, Janet, and Nathan Bronson. “Facebook's Top Open Data Problems - Meta Research.” *Meta Research*, 21 Oct. 2014, https://research.facebook.com/blog/2014/10/facebook-s-top-open-data-problems/.