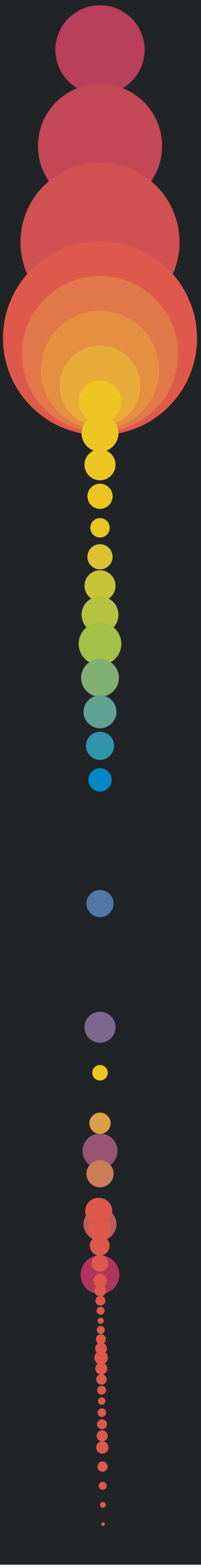
accessnow.org



DATA MINIMIZATION

KEY TO PROTECTING PRIVACY AND REDUCING HARM

accessnow

Access Now defends and extends the digital rights of users at risk around the world. By combining direct technical support, comprehensive policy engagement, global advocacy, grassroots grantmaking, legal interventions, and convenings such as Rightscon, we fight for human rights in the digital age.

This paper is an Access Now publication. It was written by Eric Null, Isedua Oribhabor, and Willmary Escoto. We would like to thank the Access Now team members who provided support, including Estelle Massé, Daniel Leufer, Gaspar Pisanu, Jennifer Brody, Sage Cheng, Juliana Castro, and Donna Wentworth.

MAY 2021

TABLE OF CONTENTS

INTRODUCTION TO DATA MINIMIZATION: WHAT IT IS AND WHY IT MATTERS

While data minimization defies any single definition, the simplest and most useful definition is that any organization (whether private company, public entity, or government body) collecting data should collect only the data necessary to provide their product or service, and nothing more. Minimizing data means collecting data only for an immediate and necessary purpose, not hoarding the data on the "off-chance that it might be useful in the future."[1] More specifically, organizations should limit 1) the scope of the data they collect, 2) the amount of data they collect within that narrow scope, and 3) the retention of that data.[2] One obvious example of the data minimization principle in action is "a data controller shall not continuously process the precise and detailed location of the vehicle for a purpose involving technical maintenance or model optimization."[3]

Data minimization is a core principle of data protection, and is part of the "Fair Information Practice Principles."[4] It is also part of an overarching concept referred to as privacy-by-design and by-default, which encourages organizations to build privacy into their products and services up-front, rather than viewing it as an afterthought. Other privacy-by-design principles include limited data retention and purpose limitation. Minimization is interconnected with these different concepts because as data is no longer necessary to fulfill an immediate and necessary purpose, organizations should no longer retain it.

Generally, organizations have not taken data minimization seriously. They have collected and retained data with impunity, and many have failed to follow safeguards or to take a disciplined approach that protects individual privacy. One study of companies in Europe showed that 72% gathered data they ended up not using.[5] Another global report showed that 55% of all data collected is "dark data" that is not used for any purpose after being collected.[6] The reasoning is clear. Organizations want as much data as possible to monetize (through, for instance, behavioral advertising), track people, or make some other use of the data in the future, potentially to train a machine learning model or sell the

[1] International Commissioners Office, *Principle (c): Data minimisation,* https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation.

[2] First, "the *possibility* to collect personal data about others should be minimized," then "within the remaining possibilities, *collecting* personal data should be minimized," and finally, "*how long* collected personal data is stored should be minimized." *Terminology for Talking about Data Minimization*, IETF (2010), https://tools.ietf.org/id/draft-hansen-privacy-terminology-00.html (emphasis added). Data retention is a closely related principle that ensures once data has served its purpose, the organization deletes the data.

[3] Commission Nationale Informatique & Libertes (NCIL), *Compliance Package - Connected Vehicles and Personal Data* (Oct. 2017), https://www.cnil.fr/sites/default/files/atoms/files/cnil_pack_vehicules_connectes_gb.pdf at 10.

[4] *Fair Information Practice Principles*, International Association of Privacy Professionals, https://iapp.org/resources/article/fair-information-practices. Note, however, that FIPPs defines collection limitations broadly: "There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject." *Id*. Our definition is stricter, focusing on what is necessary for the product or service.

[5] *Big Data's Failure: The struggles businesses face in accessing the information they need,* Pure Storage (July 2015), https://info.purestorage.com/rs/225-USM-292/images/Big%20Data%27s%20Big%20Failure_UK%281%29.pdf?aliId=64921319.

[6] Companies Collect a Lot of Data, But How Much Do They Actually Use?, Priceonomics, https://priceonomics.com/companies-collect-a-lot-of-data-but-how-much-do.

information to a data broker or the government. Data has become a commodity to many organizations, and few will change their business model in recognition of privacy being a human right.[7]

App developers, for example, have struggled with data minimization. In 2013, researchers studied 100 apps and found that 56 of them, including the “Angry Birds” and infamous “Brightest Flashlight” apps, collected geolocation information they had no business collecting.[8] The U.S. Federal Trade Commission (FTC) clamped down on the app makers after determining the apps deceived individuals and once downloaded, could tap into their calendar, location, and camera settings.[9]

More recently, the augmented reality mobile game “Pokémon GO” failed to minimize its data collection. As millions of people across the globe downloaded the app, they not only gave away access to their location and camera (somewhat necessary for the game), but also granted access to their Google accounts, which would allow Pokémon access to their photos, calendar, email, and other documents — leading critics to call the game a “huge security risk.”[10] After privacy advocates criticized the app for profiting from tracking users movements,[11] Niantic, the app’s developer, tooks steps to respond to the public backlash.[12]

Some view more expansive data collection (particularly for behavioral ads) as beneficial to both publishers and consumers.[13] However, there are reasons to doubt the extent of those benefits. First, one study showed that most of the extra value of behavioral advertising goes to ad tech companies, not the individual or the publisher.[14] Second, while most assume that behavioral advertising is superior to contextual advertising, that assumption is not a given, as contextual advertising has lacked

[7] *See* Eric Null, *Ask Apple: Facebook Doesn’t Give a Damn about Privacy Protections*, Access Now (Mar. 29, 2021), https://www.accessnow.org/facebook-apple-privacy-war (explaining that Facebook and CEO Mark Zuckerberg were furious that Apple implemented a pro-privacy iOS update because it will affect Facebook and small businesses).

[8] Bob Sullivan, *A shock in the dark: Flashlight app tracks your location*, NBC News (Jan. 16, 2013), https://www.nbcnews.com/technolog/shock-dark-flashlight-app-tracks-your-location-1B7991120.

[9] Federal Trade Commission, Press Release, *Android Flashlight App Developer Settles FTC Charges It Deceived Consumers* (Dec. 5, 2013), https://www.ftc.gov/news-events/press-releases/2013/12/android-flashlight-app-developer-settles-ftc-charges-it-deceived; Robert McMillan, *The Hidden Privacy Threat of … Flashlight Apps?*, Wired (Oct. 20, 2014), https://www.wired.com/2014/10/iphone-apps.

[10] Laura Hudson, *How to Protect Privacy While Using Pokémon Go and Other Apps*, N.Y. Times (July 12, 2016), https://www.nytimes.com/2016/07/14/technology/personaltech/how-to-protect-privacy-while-using-pokemon-go-and-other-apps.html.

[11] Yehong Zhu*, How Niantic Is Profiting Off Tracking Where You Go While Playing 'Pokémon GO,',*Forbes (July 29, 2016), https://www.forbes.com/sites/yehongzhu/2016/07/29/how-niantic-is-profiting-off-tracking-where-you-go-while-playing-pokemon-go/#1b6137a56df9.

[12] Nathan Oliverez-Giles, *‘Pokemon Go’ Creator Closes Privacy Hole but Still Collects User Data*, Wall St. Journal (July 13, 2016), https://www.wsj.com/articles/pokemon-go-creator-closes-privacy-hole-but-still-collects-user-data-1468363704.

[13] James Ewen, *What Is Behavioral Targeting? - All You Need to Know in 2020*, Tamoco (Sept. 24, 2019), https://www.tamoco.com/blog/what-is-behavioral-targeting.

[14] Veronica Marotta *et al.*, *Online Tracking and Publishers’ Revenues: An Empirical Analysis*, Workshop on the Economics of Information Security (May 2019), https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_38.pdf.

investment over the past two decades.[15] And third, people generally view being followed around the internet by the same ads as creepy, making the ads less effective.[16]

In this paper we discuss why data minimization is key to protecting privacy and explain how it can reduce the harm of data collection and exploitation. We then offer decision-makers, such as lawmakers, software developers, and others engaged in developing or implementing data minimization policies, guidance on applying data minimization principles in the context of **addressing civil rights-related harms**, **curbing the negative impact of behavioral advertising**, and **training machine learning (ML) systems**.

WHY DATA MINIMIZATION IS A HUMAN RIGHTS ISSUE

Privacy is a human right, and data minimization is a human rights issue. The most important impact of strong data minimization is harm reduction: data that is not collected cannot harm people. As organizations collect more data, the potential for and real harms to people grow. Reducing the amount of data collected is important for at least two reasons: people do not want organizations collecting every bit of information about them, and personal information can be, and often is, misused in ways that perpetuate significant harms.

People do not want organizations collecting extensive amounts of data

People generally question the wisdom of allowing organizations to collect any and all data about them. People are often unaware of how organizations use their data, and they feel they have little control over data practices.[17] A 2019 Pew survey showed 81% of U.S. respondents felt the potential risks they face because of data collection outweighs the benefits, and a majority (79%) also reported being concerned about the way their data is being used by companies.[18]

People across the world feel the same way. One study of more than 25,000 people in 40 countries showed that "7 in 10 people are concerned about sharing personal information, while two-thirds of the global population does not like the current privacy practices of most data collectors."[19] Another

[15] Becky Chao & Eric Null*, Paying for Our Privacy: What Online Business Models Should Be Off-Limits?,* Open Technology Institute (Sept. 17, 2019), https://www.newamerica.org/oti/reports/paying-our-privacy-what-online-business-models-should-be-limits ("As [DuckDuckGo General Counsel Megan Gray] stated, contextual ads have suffered from a lack of innovation.").

[16] Leslie K. John *et al.*, *Ads That Don't Overstep: How to make sure you don't take personalization too far*, Harvard Business Review Magazine (Jan.-Feb. 2018), https://hbr.org/2018/01/ads-that-dont-overstep (showing that people are less likely to engage with ads that used inferred information or information collected from third parties).

[17] *Better Machine Learning Through Data Minimization,* Privatar (Mar. 5, 2020), https://www.privitar.com/blog/better-machine-learning-through-data-minimization.

[18] Brooke Auxier *et. al., Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information,* Pew Research Center (Nov. 15, 2019), https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information.

[19] *Global Crisis In Trust Over Personal Data*, Worldwide Independent Network of Market Research (July 20, 2020), https://winmr.com/global-crisis-in-trust-over-personal-data.

Europe-based survey found that “41% [of respondents did] not want to share [specific personal] data with private companies.”[20] Further, while 72% were aware of the privacy settings on their smartphones, 24% of consumers did not know how to check the privacy settings on apps.[21] Eighty-one percent of Australians consider it a “misuse for an organization to ask for information that doesn't seem relevant to the purpose of the transaction, up 7% since 2017.”[22] And while 85% of the Australian community surveyed had a “clear understanding of why they should protect their personal information … 49% [said] they don’t know how.”[23] Unfortunately, because expansive data collection is the norm, people are unable to or have great difficulty expressing their privacy preferences, and end up resigned to using the services they want with data practices they dislike.[24]

Collecting extensive amounts of data causes extensive harm

Expansive data collection has caused significant harm, and risk of harm, for people. These harms range from the more obvious identity theft and physical harms to less obvious examples, such as relationship harms (due to loss of confidentiality), emotional or reputational harms (due to private information becoming public), or chilling effects on speech or activity (due to a loss of trust in government or other organizations).[25]

Particularly troubling are discrimination-related harms.[26] Data collection and processing can reduce opportunities for Black, Hispanic, Indigenous, and other communities of color, or actively target them for discriminatory campaigns and deception.[27] For instance, during the 2016 election, the Russian Internet Research Agency used Facebook and Twitter’s audience filters feature to target Black people to discourage them from voting.[28] Research provides ample evidence of data-driven discrimination woven into everyday life, impacting housing, employment, lending and credit, and more. For instance, a study in the U.S. found that biases in “algorithmic strategic pricing” resulted in Black and Latino borrowers paying higher interest rates on home purchase and refinance loans when compared

[20] *Your rights matter: Data protection and privacy - Fundamental Rights Survey,* European Union Agency for Fundamental Rights (June 18, 2020), https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-fundamental-rights-survey-data-protection-privacy_en.pdf, at 3.

[21] *Id*. at 7.

[22] Daniella Kafouri & John Pane, *Australian community attitudes toward privacy survey 2020,* International Association of Privacy Professionals (Dec. 3, 2020), https://iapp.org/news/a/australian-community-attitudes-towards-privacy-survey-2020.

[23] *Id*.

[24] Joseph Turow, *The Tradeoff Fallacy*, Univ. of Penn. (June 2015), https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf (“a majority of Americans are resigned to giving up their data…”).

[25] Daniel J. Solove & Danielle Keats Citron, *Privacy Harms*, George Washington School of Law (2021), https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2790&context=faculty_publications.

[26] *Id*.

[27] Cameron F. Kerry, *Federal privacy legislation should protect civil rights*, Brookings Institute (July 16, 2020), https://www.brookings.edu/blog/techtank/2020/07/16/federal-privacy-legislation-should-protect-civil-rights.

[28] Scott Shane & Sheera Frenkel, *Russian 2016 influence operation targeted African-Americans on social media,* The New York Times Magazine (Dec. 17, 2018), https://www.nytimes.com/2018/12/17/us/politics/russia-2016-influence-campaign.html; Jack Stubbs, *Facebook says Russian influence campaign targeted left-wing voters in U.S., UK*, Reuters (Sept. 15, 2020), https://www.reuters.com/article/usa-election-facebook-russia/facebook-says-russian-influence-campaign-targeted-left-wing-voters-in-u-s-uk-idUSKBN25S5UC; *Report of the Select Committee on Intelligence in the U.S. Senate on Russian Active Measures Campaigns and Interference in the 2016 Election, Volume 2: Russia’s Use of Social Media with Additional Views*, at 35, https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf (a statement from Renee DiResta, Director of Research at New Knowledge, a cybersecurity company, indicated “Voter suppression narratives were … specifically targeting the Black audiences.”).

to White and Asian borrowers, costing Black and Latino borrowers $250-500 million every year.[29] Another example is companies using Facebook's ad filters to exclude certain users from seeing job ads, such as Uber and NTB Trucking displaying jobs only to men,[30] and 40 other companies categorically excluding older workers from seeing their ads.[31] In the financial services industry, several companies violated Facebook's anti-discrimination policy when it targeted ads ranging from credit cards to home equity loans to restricted groups.[32] U.S. companies Staples, Home Depot, Discover Financial Services, and Rosetta Stone used data on users' physical location to display higher online prices and fewer deals for people in low-income neighborhoods.[33] This kind of data should not be collected in the first instance unless it is necessary to provide the service, or as we argue below, to audit data processing systems for bias. Once collected, it certainly should not be used to discriminate.

Data minimization reduces harm by limiting surveillance and increasing security

Another risk of extensive data collection is use of the information for government surveillance, which can lead to abuse of government authority and chilling effects on free expression. Data minimization can reduce those harms as well. When a government seeks information from a company like Signal, which offers end-to-end encryption of all communications (preventing any third party from viewing the contents of those communications) and keeps the data they collect about users to the bare minimum, the company has no information to give those authorities. When the U.S. government recently made such a request, including asking for the names and addresses of users, Signal's response was that it "couldn't provide any of that. It's impossible to turn over data we never had access to in the first place."[34] If more companies adopted this kind of robust data minimization, fewer people would be subject to privacy violations, government surveillance, and abuse.

Data minimization is also an important component of data security. As the unnecessary collection and retention of data increases, the growing treasure trove of data becomes a target for third parties, whether it is law enforcement or malicious hackers. Amazon's most recent transparency report, which covers Amazon's shopping site as well as its Echo, Ring, and Fire products, shows an 800% increase in

[29] Robert Bartlett *et al*., *Consumer Lending Discrimination in the FinTech Era*., Univ. of California School of Law (2017), http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf?_ga=2.236934529.1452837941.1619188513-1084532383.1619188513; Laura Counts, *Minority homebuyers face widespread statistical lending discrimination, study finds*, Univ. of California Berkeley Haas School of Business (Nov. 13, 2018), https://newsroom.haas.berkeley.edu/minority-homebuyers-face-widespread-statistical-lending-discrimination-study-finds.

[30] Ariana Tobin & Jeremy B. Merrill, *Facebook Is Letting Job Advertisers Target Only Men*, ProPublica (Sept. 18, 2018), https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men.

[31] Jeff Larson *et al*., *These Are the Job Ads You Can't See on Facebook If You're Older,* ProPublica (Dec. 19, 2017), https://projects.propublica.org/graphics/facebook-job-ads.

[32] Corin Faife & Alfred Ng*, Credit Card Ads Were Targeted by Age, Violating Facebook's Anti-Discrimination Policy,* Markup (April 29, 2021), https://themarkup.org/citizen-browser/2021/04/29/credit-card-ads-were-targeted-by-age-violating-facebooks-anti-discrimination-policy.

[33] Jennifer Valentino-DeVries *et al*., *Websites Vary Prices, Deals Based on Users' Information*, Wall St. Journal (Dec. 24, 2012), https://www.wsj.com/articles/SB10001424127887323777204578189391813881534.

[34] *Grand jury subpoena for Signal user data, Central District of California,* Signal (Apr. 27, 2021), https://signal.org/bigbrother/central-california-grand-jury.

law enforcement requests for user data in 2020 alone.[35] The spike is likely related to how much data Amazon holds about its users.

The harm caused by data breaches, hacks, or unauthorized access of data within an organization is simply too great to justify collecting more data than is necessary to provide a product or service. Organizations have the responsibility to secure and protect the data they process. Minimizing the amount of data they collect is one of the best, most human rights-respecting ways to prevent privacy violations and harms.

DATA MINIMIZATION IN PRACTICE: TRICKY USE CASES

While data minimization is straightforward in principle, the details are complicated. Below we explore how data minimization principles should be applied in the context of addressing civil rights-related harms, curbing the negative impact of behavioral advertising, and improving machine learning (ML) systems.

Allow organizations to collect data for the purpose of tracking civil rights violations and benefiting underrepresented populations

Organizations should frequently audit their systems to ensure they are limiting the data they collect to that which is necessary to provide their service, and thus limit the harm they could potentially cause. Indeed, if they do not, they may run afoul of their nation's laws.[36]

Data minimization does not preclude the collection of data such as race and gender when it is necessary to provide a service. But one could argue that it prevents organizations from collecting information on race, gender, or other protected attributes (1) to study whether the organization violates, or facilitates the violation of, civil rights laws and (2) to benefit underrepresented populations, such as communities of color. Yet without such information, an organization (and outside auditors and watchdogs) would have a difficult time determining whether its practices fail to protect civil rights.

In that case, we recommend carving out a narrow exception to strict data minimization requirements for these purposes.[37] Organizations should be allowed to collect data on protected classes where their purpose is to address their own discriminatory practices and mitigate or eliminate the harms or to benefit certain underrepresented populations.

[35] Zach Whittaker, *Amazon says government demands for user data spiked by 800% in 2021*, Tech Crunch (Feb. 1, 2021), https://techcrunch.com/2021/02/01/amazon-government-demands-spiked.

[36] The United States Federal Trade Commission recently stated that biased machine learning violates the law. Elisa Jillson, *Aiming for truth, fairness, and equity in your company's use of AI*, Federal Trade Commission (Apr. 19, 2021), https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

[37] Any attempts to address or "fix" systems that have been determined to be biased should still operate under a strict data minimization principle, and should not allow for expansive collection of data on protected class.

RECOMMENDATION

In the context of a strong data protection framework, allow organizations to collect data on protected classes for the purpose of civil rights auditing or to benefit underrepresented populations.

Comprehensive privacy legislation could include strong data minimization requirements with a specific exception for "the purpose of advertising, marketing, or soliciting economic opportunities to underrepresented populations in a fair, non-deceptive, and non-predatory manner" and "legitimate internal testing for the purpose of preventing unlawful discrimination or otherwise determining the extent or effectiveness of the [organization's] compliance" with civil rights laws.[38] With those exceptions, organizations could audit their systems for bias but still be subject to data minimization requirements.

Of course, once that data on protected classes is collected and stored, it should be put to no other use, and should be strictly protected against unauthorized access, unauthorized disclosure, and other data protection violations. There are too many examples of civil rights harms from automated technology, such as Facebook's allowing advertisers to target audiences based on protected categories,[39] and Facebook's look-alike audiences recreating bias in the datasets its advertisers provide.[40] No organization should build systems that discriminate based on protected class.

Reduce the harm of behavioral advertising by limiting the collection of data for advertising purposes

Behavioral advertising is the dominant online business model. It is generally defined as targeting an advertisement to individuals based on their past behavior. This targeting entails tracking people's online activities, most often their web browsing history, app usage, or other attributes.[41] It also entails "profiling,"[42] which is defined in the EU as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person," specifically a person's preferences, behavior, or location.[43]

People have criticized the behavioral advertising business model for many reasons, including because it is invasive and creepy, it can lead to the profiling and targeting of users, and it increases the risks of

[38] *See The Online Civil Rights and Privacy Act of 2019*, Free Press and Lawyers' Committee for Civil Rights Under Law, https://www.freepress.net/sites/default/files/2019-03/online_civil_rights_and_privacy_act_of_2019.pdf, at Section 3(g).
[39] Julia Angwin *et al.*, *Facebook (Still) Letting Housing Advertisers Exclude Users by Race*, ProPublica (Nov. 21, 2017), https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.
[40] Muhammad Ali *et al.*, *Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes*, https://arxiv.org/pdf/1904.02095.pdf.
[41] The state of California has defined it as the "targeting of advertising to a consumer based on the consumer's personal information obtained from the consumer's activity across businesses, distinctly branded websites, applications or services, other than the business, distinctly-branded website, application, or service with which the consumer intentionally interacts." CPRA §1798.140(k), https://thecpra.org.
[42] Francesco Banterle, *Early thoughts on behavioral advertising and the GDPR: a matter of discrimination?*, IP Lens (Sept. 19, 2017), https://iplens.org/2017/09/19/early-thoughts-on-behavioral-advertising-and-the-gdpr-a-matter-of-discrimination.
[43] GDPR Article 4.4, https://gdpr-info.eu/art-4-gdpr.

discrimination.[44] As the case against behavioral advertising grows,[45] the movement to ban it grows.[46] There is at least one bill in the U.S., Senator Brown's Data Accountability and Transparency Act of 2020, that would do just that.[47] In countries that already have a data minimization requirement, such as in the EU under the General Data Protection Regulation, behavioral advertising may already be significantly curtailed due to the strong opt-in provisions in the law.[48]

While the evidence of harm is abundant, there is little to show the benefit of behavioral advertising for the companies deploying it. It may not be as effective as claimed. A recent study found that publishers retain only 4% of the increased revenue from behavioral advertising.[49] In 2019, when *The New York Times* cut off ad exchanges and turned to contextual advertising, it saw its revenues rise.[50] Data from the Dutch broadcaster NPO showed that when it ditched behavioral advertising for contextual ads across its sites for the first half of 2020, its revenue increased each month.[51] There are also costs associated with the collection, retention, and maintenance of all the data collected for advertising.[52] Rather than invest in crafting privacy-protective alternatives, "[m]uch of the innovation [in ads] has gone specifically into matching people to ads based on their behavior rather than the context of the current website."[53]

Needless to say, behavioral advertising presents a problem for data minimization. How does an organization minimize the data it collects when it uses data to feed its behavioral advertising system? If behavioral advertising is an organization's primary source of revenue, the organization may believe that it "needs" any and all data about its users to ensure that it delivers the most relevant ads. Notably, it appears these same organizations rarely explore whether different, less intrusive, data processing can meet their objectives, as they should to comply with principles of necessity and proportionality in the EU.[54] If challenged on the basis of violating data minimization principles, the developers of the

[44] *Chao & Null,* Paying for Privacy, https://www.newamerica.org/oti/reports/paying-our-privacy-what-online-business-models-should-be-limits.

[45] Natasha Lomas, *The case against behavioral advertising is stacking up,* Tech Crunch (Jan. 20, 2019), https://techcrunch.com/2019/01/20/dont-be-creepy; Gilad Edelman, *Why Don't We Just Ban Targeted Advertising?* Wired Magazine (Mar. 22, 2020), https://www.wired.com/story/why-dont-we-just-ban-targeted-advertising.

[46] Ban Surveillance Advertising, https://www.bansurveillanceadvertising.com.

[47] Data Accountability and Transparency Act of 2020, https://www.banking.senate.gov/imo/media/doc/Brown%20-%20DATA%202020%20Discussion%20Draft.pdf.

[48] Francesco Banterle, *Early thoughts on behavioral advertising and the GDPR: a matter of discrimination?*, IP Lens (Sept. 19, 2017), https://iplens.org/2017/09/19/early-thoughts-on-behavioral-advertising-and-the-gdpr-a-matter-of-discrimination.

[49] Natasha Lomas*, Targeted ads offer little extra value for online publishers, study suggests*, Tech Crunch (May 31, 2019), https://techcrunch.com/2019/05/31/targeted-ads-offer-little-extra-value-for-online-publishers-study-suggests.

[50] Jessica Davies*, After GDPR, The New York Times cut off ad exchanges in Europe — and kept growing ad revenue*, Digiday (Jan. 16, 2019), https://digiday.com/media/gumgumtest-new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue. While not every publisher is *The New York Times*, it indicates a potential path forward with contextual advertising.

[51] Natasha Lomas, *Data from Dutch public broadcaster shows the value of ditching creepy ads*, Tech Crunch (July 24, 2020), https://techcrunch.com/2020/07/24/data-from-dutch-public-broadcaster-shows-the-value-of-ditching-creepy-ads/?guccounter=1.

[52] Christopher Tozzi, *5 hidden costs of big data*, Precisely (June 8, 2020), https://www.precisely.com/blog/big-data/the-hidden-costs-of-big-data.

[53] *Chao & Null,* Paying for Privacy at 12, https://www.newamerica.org/oti/reports/paying-our-privacy-what-online-business-models-should-be-limits/legislation-could-promote-privacy-protective-business-models (quote from DuckDuckGo's General Counsel Megan Gray). *See also* Gabriel Weinberg, *What if we all just sold non-creepy advertising?*, N.Y. Times (June 19, 2019), https://www.nytimes.com/2019/06/19/opinion/facebook-google-privacy.html.

[54] Necessity & Proportionality, European Data Protection Supervisor, https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en.

"Brightest Flashlight" app might argue that they collected geolocation data to serve location-based ads as a source of revenue, and that could hypothetically satisfy a weak or non-existent data minimization requirement, leaving aside the transparency issues that would remain. Facebook allows advertisers of products or services to target any of its users or group of users, which means the company can and generally does collect seemingly unlimited amounts of data about its users.

Short of banning behavioral advertising, which would provide significant privacy benefits, we believe that regulators should at minimum intervene to ensure that organizations reduce the harms that result from behavioral advertising. Data minimization and retention limits can play an important role in reducing these harms.

RECOMMENDATION

Regulators that do not ban behavioral advertising should at minimum require limits to the data collected for this purpose.

Specifically, an organization that collects data for advertising purposes—which should already be minimized to follow general data minimization principles—should be required to delete, not merely de-identify, that information, as well as any information they inferred from that data, after 30 days. Old data about a person's "interests" based on their browsing history can not only miss the mark but will also be stale, and users are likely to perceive ads based on ancient browsing history as "creepy" and intrusive. People are less likely to engage with ads they view as creepy.[55] These limitations will not eliminate, but can potentially reduce, the harms associated with indefinite retention of data and profiling users.

Use thoughtful data minimization to build better machine learning systems

The training of machine learning (ML) systems can require extremely large datasets.[56] The complexity of some approaches to machine learning can make it difficult to understand whether they operate under a data minimization principle.[57]

ML systems are better when trained on *good* data, rather than simply as much data as possible.[58] There is a misguided assumption that more data is always better, and that all data is useful. However, those building ML systems should take into account data minimization principles because inappropriate data can negatively impact the performance of a system in problematic ways. There are ways to build

[55] Leslie K. John *et al.*, *Ads That Don't Overstep: How to make sure you don't take personalization too far*, Harvard Business Review Magazine (Jan.-Feb. 2018), https://hbr.org/2018/01/ads-that-dont-overstep (showing that people are less likely to engage with ads that used inferred information or information collected from third parties).

[56] Daniel Leufer *et al.*, *AI Myths - AI can solve any problem,* https://www.aimyths.org/ai-can-solve-any-problem.

[57] Abigail Goldsteen *et al.*, *Data Minimization for GDPR Compliance in Machine Learning Models*, Cornell University (2020), https://arxiv.org/pdf/2008.04113.pdf.

[58] *See generally* Eliza Strickland, *OpenAI's GPT-3 Speaks! (Kindly Disregard Toxic Language*), IEEE Spectrum (Feb. 1, 2021), https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/open-ais-powerful-text-generating-tool-is-ready-for-business.

privacy-preserving techniques like minimization into the various phases of machine learning.[59] These techniques allow ML architects to be intentional, careful, and selective with the data they collect, ensuring that they both minimize privacy risks and build an effective, functional system.

Thus, developers should try to ensure the data collected will help the ML system perform well and that it is collected in a human-rights compliant and ethical manner.[60] One of the most advanced achievements of ML to date has been OpenAI's language generation system, GPT3, which has produced remarkable examples of "natural-sounding" language. However, partly because it was trained on an enormous database of text scraped from all corners of the internet, including Reddit, it has been consistently shown to reproduce racist, sexist, and otherwise problematic language. While "big data" has enabled impressive performance in some tasks, it has also burdened the model with deeply troubling prejudices.[61]

RECOMMENDATION

Machine learning developers should adopt a method to perform data minimization for ML models that minimizes the effects on model performance and safeguards privacy rights.

Data minimization does not mean that developers cannot collect certain kinds of data. Rather, it means that organizations should collect only what is necessary for their service or product. For example, a ML tool designed to increase racial and gender diversity in a hiring process may need to collect demographic information on race and gender. Developers should also be permitted to collect this kind of information to identify or audit areas of bias in ML systems, as discussed above.

[59] *Better Machine Learning Through Data Minimization,* Privatar (March 5, 2020), https://www.privitar.com/blog/better-machine-learning-through-data-minimization.
[60] Bernard Marr, *Why AI Would Be Nothing Without Big Data*, Forbes (June 9, 2017), https://www.forbes.com/sites/bernardmarr/2017/06/09/why-ai-would-be-nothing-without-big-data.
[61] Strickland, *OpenAI's GPT-3 Speaks!*.

CONCLUSION AND RECOMMENDATIONS

Data minimization is key to protecting privacy and reducing privacy-related harms. Data not collected cannot be used to harm people. Without minimizing data collected, privacy harms will continue to compound.

We recommend the following:

- ➔ **In the context of a strong data protection framework, allow organizations to collect data on protected classes for the purpose of civil rights auditing or to benefit underrepresented populations**: Organizations should be allowed to collect data on protected classes where their purpose is to address their own discriminatory practices and mitigate or eliminate the harms or to benefit certain underrepresented populations.

- ➔ **Regulators that do not ban behavioral advertising should at minimum require limits to the data collected for this purpose:** An organization that collects data for advertising purposes—which should already be minimized to follow general data minimization principles—should be required to delete, not merely de-identify, that information, as well as any information they inferred from that data, after 30 days.

- ➔ **Machine learning developers should adopt a method to perform data minimization for ML models that minimizes the effects on model performance and safeguards privacy rights**: Developers should try to ensure the data collected will help the ML system perform well and that it is collected in a human-rights compliant and ethical manner.

For more information, contact:

Access Now Data Protection Team | dataprotection@accessnow.org
Media Requests | press@accessnow.org