VIRGINIA JOURNAL OF LAW & TECHNOLOGY

FALL 2022 UNIVERSITY OF VIRGINIA VOL. 26, NO. 5

Data Inferences and Free Speech

Anqi Wang[*]
Han Liu[*]

[*] MSc in Social Science of the Internet, Oxford Internet Institute at the University of Oxford.

[*] Associate Professor, Tsinghua University School of Law; JSD, Yale University.

Abstract

Data inference is a form of speech. But free speech should not be exercised as a natural right to protect all types of data inferences. For decades, privacy scholars have argued against this data aggregation and resulting inferences from a perspective of privacy and reputation. This paper argues that instead of taking a privacy-first approach, data inference critics should set their agenda focusing on certain kinds of harmful data inferences, because those digital equivalents of traditionally unprotected speech may remain outside the scope of the First Amendment. Guided by Erica Goldberg's free speech consequentialism, this article designs a framework of spectrum to illustrate how free speech can be applicable to different types of data speech or not. It posits the individual speech category which represents self-determination and individual liberty should sit on the end of spectrum of absolute free speech protection. The devil data speech category, located at the other end of spectrum, includes harmful data speech that poses imminent danger to others and should not enjoy free speech protections.

For the categories of data inferences sitting in the middle of the spectrum, which are mostly commercial data inferences, this article offers a three-dimensional assessment procedure with a listener-centric approach. It argues that data subjects, who are also the listeners in the data speech context when we see data inferences as speakers, are also entitled to a right to free speech which diminishes with the unbridled expansion of speakers' freedom of expression. Nevertheless, this negative correlation is overlooked by data aggregators, governments, and even the listeners themselves. This article aims to bring clarity to the relationship between data inferences and the free speech protection, and offer potential pathways for regaining digital free speech rights for data subjects.

Table of Contents

I. INTRODUCTION

Data only realizes its value when aggregated en mass. Sometimes manifested as customized speech,[1] data inferences,[2] or code speech,[3] data aggregation refers to a method where data processors generate new information based on primary data created by users. Then, through technological advances in data mining, raw forms of personal data are assembled and transformed into new dossiers of human capital. This is furthered by the growth of digital tracing and facilitated by expanding data storage capacity. Digital aggregation, in turn, provides a strong engine for growing digital capitalism.[4]

The social consequences of aggregated data, however, are not as predictable as their operating mechanisms. When digital surveillance capitalism encourages data aggregation, the economic incentive becomes tempting enough to sometimes challenge social norms that human society has long valued such as privacy, reputation, and social equity.[5] Unsurprisingly, the spread of harmful data inferences has prompted some scholars to seek to restrict such data inferences.[6]

[1] Daniel Rauch, *Customized Speech and the First Amendment*, 35 HARV. J. L. & TECH. 405 (2022).

[2] Sandra Wachter & B. D. Mittelstadt, *A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI*, 494 COLUM. BUS. L. REV. (2019).

[3] Andrea M. Matwyshyn, *Hacking Speech: Informational Speech and the First Amendment*, 107 NW. U. L. REV. 795 (2013).

[4] Kenneth Cukier & Viktor Mayer-Schoenberger, *The Rise of Big Data: How it's Changing the Way We Think about the World,* 92 FOREIGN AFFAIRS 28, 29 (2013).

[5] Shoshana Zuboff, *Surveillance capitalism and the challenge of collective action*, 28 NEW LABOR FORUM 10–29 (2019).

[6] Wachter & Mittelstadt, *supra* note 2, at 495 (proposing a right to reasonable data inferences for closing the gap posed by "high risk inferences" that damage privacy or reputation or have low verifiability.); Mary Anne Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892 (2019) (arguing in response to the marketplace of ideas theory of the First Amendment, "[w]hatever merit these claims may have had in the past, they cannot be sustained in the digital age.").

On the other side of the debate are scholars and courts who believe that data aggregation and their digital inferences are protected by the First Amendment. In the judicial arena, for example, a web-scraping company, hiQ, argued that the information it scraped from LinkedIn was protected as a form of free expression.[7] The Northern District of California agreed and held that the inferences were protected by the California state constitution.8 Some saw this ruling as a victory of a marketplace of ideas and free speech, offering protections to public data scraping.[9]

Scholars have argued that a right to scrape and aggregate public data provides more room to journalists and researchers who conduct investigative work for the public interest.[10] And even if there are harmful data inferences, they should be weighed against the countervailing benefits.[11] In response to privacy challenges,[12] some have suggested that the barriers privacy laws pose to the sale and disclosure of personal information are unconstitutional under

[7] *hiQ Labs, Inc. v. LinkedIn Corp*., 938 F.3d 985, 992 (9th Cir. 2019).

[8] *hiQ Labs, Inc. v. Linkedin Corp.*, 485 F. Supp. 3d 1137, 1142 (N.D. Cal. 2020).

[9] Camille Fischer & Andrew Crocker, *Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data,* ELECTRONIC FRONTIER FOUNDATION (Sept. 10, 2019), https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.

[10] Jacquellena Carrero, *Access Granted: A First Amendment Theory of Reform of the CFAA Access Provision*, 120 COLUM. L. REV. 134-35 (2020). Komal S. Patel, *Testing the Limits of the First Amendment: How Online Civil Rights Testing Is Protected Speech Activity*, 118 COLUM. L. REV. 1473, 1501-03 (2018).

[11] Rauch, *supra* note 1, at 452 (highlighting three countervailing benefits: informing and engaging voters, empowering the marginalized, and checking government overreach).

[12] Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373, 1376 (2000). (Suggesting that seeing the collection and exchange of personally-identified data as "speech" is essentialist. This argument "equates the market exchange of information for value with the highest sort of protected expression, and thus ignores that the relation between personally-identified information and expression is far more complex, and far less direct.")

the First Amendment law because it is difficult to demonstrate that personal information is merely a matter of private concern.[13]

We do not support this assertion. Data can be considered speech, but this cannot lead to the conclusion that data inferences merit full-scale free speech protection. Even data speech covered by the First Amendment does not always receive free speech protection.[14] Currently, there is a fierce debate among scholars over whether data is entitled to free speech protections, which focuses on three waves of viewpoints: some deny data as a form of speech, and some argue all data speech should be protected by free speech rights, and others that argue free speech claims should be granted to only limited types of data inferences.[15] The third view sees excluding data inferences from speech as risky, but this is not to say free speech must protect harmful data speech.[16]

This article, in favor of the viewpoint of the third wave, argues that the conventional application of free speech protection does not easily stretch to accommodate all types of data aggregators. First, current discussions ignore the fact that not only are data inferences a form of free speech, but users or listeners also have a free-speech right to generate data and create content. As users become less inclined to disclose information when they are aware of the surveillance,[17] data aggregators' free speech rights conflict with listeners' free speech rights.

Second, the types of speech that are traditionally not given the full measure of protection include fighting words, obscenity,

[13] Eugene Volokh, *Freedom of Speech and Information Privacy: The Troubling Implications of A Right to Stop People from Speaking About You*, 52 STAN L. REV. 104 (2000).
[14] Frederick Schauer, *Out of Range: On Patently Uncovered Speech*, 128 HARV. L. REV. F. 346, 347-348 (2015).
[15] *See* discussions *infra* Section II. A.
[16] Leslie Kendrick, *Must Free Speech Be Harmful?*, 2020 U. CHI. LEGAL F. 105 (2020). (Suggesting that "Protection of harmful conduct is not a necessary feature of any right, including a free speech right.")
[17] Jonathon Penney, *Internet Surveillance, Regulation, and Chilling Effects Online: a Comparative Case Study*, 6 INT. P. R. (2017).

libel, incitement, child pornography and commercial speech in some contexts.[18] These expressions also widely exist in aggregated data.[19] But free speech advocates have failed to acknowledge the digital parallels of traditionally unprotected speech in the realm of data inferences. If free speech rights can legitimize these harmful data inferences that have traditionally fallen outside of the First Amendment's reach, then the scope of free speech rights is actually expanded by a technological, not a legal, change of the development of data inferences.

The crucial issue that requires an answer from the third-wave argument is how to distinguish types of unprotected data inference speech from protected data inference speech. Although many types of speech are traditionally subject to constitutional review, courts and scholars have yet to articulate a coherent theory to inform the scope of the First Amendment,[20] let alone in the information technology context. The criterion to decide which data speech are harmful enough to be excluded from the protection of free speech rights requires rigorous analysis.

This article borrows from Erica Goldberg's improved version of Free Speech Consequentialism, which argues that only speech harm that analogizes conduct harm and creates an imminent and tangible danger should be precluded from the First Amendment protection.[21] It highlights that whether data inferences can be protected by free speech rights is not just a need to balance between free speech and privacy, as potential risks that data inferences trigger, once they are recognized as free speech, go beyond the realm of privacy. Instead of taking a privacy-first approach, data inference critics should set their agenda around harmful data inferences, those digital substitutions of traditionally unprotected

[18] David S. Boyce, *Commercial Speech: First Amendment Protection Clarified*, 28 U. FLA. L. REV. 610, 613 (1976).
[19] *See* discussions *infra* Part II. A.a.
[20] Toni M. Massaro, Helen Norton & Margot E. Kaminski, *SIRI-OUSLY 2.0: What Artificial Intelligence Reveals about the First Amendment*, 101 MINN. L. REV. 2481, 2487 (2017).
[21] Erica Goldberg, *Free Speech Consequentialism*, 116 COLUM. L. REV. 687, 689 (2016).

speech that remain outside the scope of the First Amendment, such as conduct-related harm or discrimination-related speech.

Based on the categorical approach affirmed by Goldberg's Free Speech Consequentialism,[22] this paper designed a framework of a spectrum to illustrate how free speech can be applicable to different types of data speech. It firstly classifies data inferences into four categories: individual speech, devil data speech, political data speech, and commercial data speech. It posits that the individual speech category, which represents self-determination and individual liberty— even data speech that is trivial and useless — should be situated on the end of the spectrum representing an absolute free speech. The devil data speech category is located at the other end of the spectrum, which includes harmful data speech that presents an imminent danger to others, and therefore should not enjoy free speech protections.

For devil data speech, we argue that whether free speech applies to data inferences does not depend on the technology itself, i.e. whether they are deepfakes, or data scrapping techniques, or other forms of algorithmic technologies that technically designed to be neutral. Instead, we propose to follow a harm-based approach for free speech protection considerations: under the harm-based approach, the level of potential conduct-related harms proportim mmonates the extent for data speech to receive free speech protections. That being said, if certain applications of deepfakes are not included under the scope of free speech, it is because their data inferences risks enough harms to speech receivers, rather than because they belong to the category of deepfakes, and any data inferences generated by deepfake technologies should be excluded from free speech protection. Similarly, if certain political data speech such as Cambridge Analytica situates outside of free speech protections, it is not because all political data inferences are excluded from free speech protection. Rather, it is the use of deceptive manipulative marketing tactics, an unprotected marketing strategy for commercial speech, that failed political data inferences to be considered as free speech. Many cases of political data

[22] *Id.* at 703-705.

speech also qualify as commercial data speech as they attached commercialized properties and therefore should be scrutinized as commercial data speech.

Although it does not aim to make an exhaustive list of types of harmful data inferences, this article exemplifies three types of harmful speech that are too devil to be included in the First Amendment protection: the deepfaked pornography, commercial data speech with deceptive marketing techniques, and the manipulative data speech sometimes reflected in political data speech. These three types of data speech run contrary to listeners' interests and have obvious defects that cannot be offset by their merits (if any), and therefore should not be considered for free speech protection.

A listener-centric approach is proposed in this article. Listeners, equally important agents as speakers in receiving speech, have interests in data inferences which are greatly compromised by excessive, harmful data speech. Tim Wu argued that preserving the basic values of the First Amendment includes providing basic protections for both speakers and listeners as a constitutional duty.[23] However, we are at the stage wherein mostly free speech rights of speakers as data aggregators, as their speech is more explicit and avid, are being emphasized as data speech.[24] There is a lack of attention to interests of listeners as data agents, including how listeners' valuable time is exploited by speakers.

It is worth noting that the category of devil data speech does not include speech-related harm that poses risks short of clear and present danger, such as privacy and reputation damage. Does this mean that listeners may only face a helpless situation when

[23] Tim Wu, *Is the First Amendment Obsolete*, 117 MICH. L. REV. 569 (2018).

[24] Especially from the perspective of speakers' self-fulfillment, speech does not have to be instrumentally beneficial for audience but instead they could be "valuable in and of itself to the free speaker". *See*, RonNell Andersen Jones, *Press Speakers and the First Amendment Rights of Listeners*, 90 U. COLO. L. REV. 499, 502 (2019); Burt Neuborne, MADISON'S MUSIC: ON READING THE FIRST AMENDMENT 98-99 (2015) (arguing that "when the interests of speakers and hearers diverge, the edge usually goes to speakers").

data inferences create emotional distress and reputation damage? To address this question, this article proposes that a three-dimensional analysis with a listener-centric approach should be carried out to assess the adequacy of free speech claims of all speech dwelling in the middle of the spectrum, mostly commercial data speech. Data inferences generalized by the unauthorized access of data controllers out of financial incentives may not enjoy free speech protections unless they 1) show public interest, 2) can be counterargued, or 3) obtain users' consent to aggregate their data. It points out that, as important digital players, users are entitled to authorize data inferences but have been muted during the legal battle between LinkedIn and hiQ. Through the evaluations of three-dimensional analysis, it proposes the right to authorize data inferences as a potential solution to curbing the abuse of freedom of speech for data generalization, especially those for commercial purposes.

This article proceeds as follows. Section II introduces three waves of views that navigate the relationship between data and free speech protections and defends the third viewpoint by analyzing why free speech does not work for harmful data inferences. Section III posits a spectrum that constitutes four types of data speech and illustrates why data inferences generated by individuals can be seen as protected data speech but devil data inferences cannot. Section IV proposes a listener-centric approach as a potential pathway for privacy advocators to incorporate data subjects ' free speech rights into their efforts to advance data protection.

II. THREE WAVES OF UNDERSTANDING ON DATA AS SPEECH

Upon close scrutiny, data inferences are such a broad category that not a single type of data inference, whether it is political speech, discriminative algorithmic predictions, or deepfake videos, has the capacity to draw an overarching conclusion as to whether free speech protects data inferences. To support this argument, we lay out three representative arguments on the relationship between data inferences and free speech. We argue that data is a kind of speech, but this cannot lead to the conclusion that all data inferences fall under constitutional protections. To exclude those data inferences that cannot be within the scope of the First Amendment, a classification of data inferences is necessary.

A. Generalized Data as a Form of Speech

There are scholarly views on whether data inferences can be seen as protected by free speech rights. There is a developing tendency to include data as speech. Nevertheless, free speech advocates have failed to explain how "data inferences as speech" can be translated into "all data inferences deserve free speech protections."

i. First Wave: Data is not Speech

Seeing data as a form of speech is not new.[25] In *Sorrell v. IMS Health Inc.* in 2011, Justice Kennedy put it explicitly that "the creation and dissemination of information are speech within the meaning of the First Amendment."[26] Privacy scholars are the major force refusing the idea of seeing data as speech and affording it free speech rights. A host of articles written by privacy scholars have investigated whether the nature of data speech can be protected by free speech rights.[27] Because the First Amendment generally protects expression unless that expression falls outside the

[25] Jane Bambauer, *Is Data Speech*?, 66 STAN. L. REV. 57 (2014); Cohen, *supra* note 12, at 1375 (suggesting that firms argue their collection and analyzation of personally-identifiable data is constitutionally-protected speech because it is information).

[26] *Sorrell v. IMS Health*, 564 U.S. 552, 567 (2011). (holding that Vermont's Prescription Confidentiality Law, which forbids prescription records from being sold or used for marketing purposes without doctors' 'consent, was unconstitutional because records are "speech in aid of pharmaceutical marketing is a form of expression protected by the Free Speech Clause of the First Amendment.").

[27] Cohen, *supra* note 12, at 1376 (arguing that taking personally-identifiable data as constitutionally-protected speech ignores "the relation between personally-identified information and expression is far more complex, and far less direct."); Daniel J. Solove, *The Virtues of Knowing Less: Justifying Privacy Protections Against Disclosure*, 53 DUKE L. J. 967, 987 (2003).

category of speech,[28] a common strategy that privacy scholars have used is to delegitimize data as simply "private speech",[29] or not speech.[30] Nevertheless, the denial of data speech seems critically weak when the argument is motivated by fear of data speech's consequences rather than the nature of data speech itself.

Furthermore, privacy scholars have turned their attention to the judicial interpretation of free speech rights; they argue that the First Amendment includes the value of privacy.[31] For example, Carrero cited *Griswold*, wherein Justice Douglas noted a constitutional right of privacy to protect intimate relations of married couples regarding reproductive rights to prove that privacy values are embodied in the First Amendment.[32] While these scholars have successfully identified the conflicting interests caused by free speech regarding privacy, their arguments are limited to findings within the framework of privacy. In a country where privacy-invading speech that damages a rape victim's reputation can be published with impunity under the First Amendment,[33] privacy-only arguments against America's exceptional commitment to free

[28] Derek E. Bambauer, *Exposed*, 98 MINN L. REV. 2025, 2091 (2014).

[29] Neil M. Richards, *Reconciling Data Privacy and the First Amendment,* 52 UCLA L. REV. 1149, 1169, 1173 (2005) (arguing that personal data is not speech because data is more commodity than expressive ideas); Ashutosh Bhagwat, *Sorrell v. IMS Health: Details, Detailing, and the Death of Privacy*, 36 VT. L. REV. 855, 856, 875-76 (2012) (admitting some grounds to treat data as speech, but ultimately advising to characterize data as less valuable, purely private speech).

[30] Helen Norton, *Manipulation and the First Amendment*, 30 WM. & MARY BILL RTS. J. 223 (2021).
(describing speech that does things rather than say things falls outside of the First Amendment's protection).

[31] Carrero, *supra* note 10, at 159-160 (summarizing scholarship on free speech and privacy to illustrate "privacy is key to First Amendment values of autonomy, thought formation, and self-governance.").

[32] Id. at 159.

[33] *Cox Broad. v. Cohn*, 420 U.S. 469 (1975). In *Cox Broadcasting*, the Court found that where a rape victim's name is publicly disclosed in a court record, the media holds a free speech right and its subsequent publication on the victim's name cannot be banned.

speech protection[34] are doomed to fail.

ii. Second Wave: All categories of Data Speech Enjoy Free Speech Protection

By contrast, free speech supporters take a firm stance on the argument that data should be seen as speech and are therefore under the First Amendment protection.[35] In assessing how courts incorporate the speech at issue into First Amendment jurisprudence, a common strategy employed by free speech supporters is a two-step method: they attempt to legitimize free speech protections to data speech at the data collection step and the new data creation step.

For the data collection step, free speech supporters attempt to prove that data collection conduct, especially those visible in public spaces, is a form of speech with a "right-to-record" jurisprudence.[36] On this view, as the data scraping behavior can be interpreted as a kind of free expression, data collectors should not be constrained by the government to gather information with their data scraping techniques. Through understanding information collection as a right to access knowledge,[37] it has been argued that the recording of information may advance democratic self-governance and the search for truth on a doctrinal level.[38] However, for automated data collection and commercial data scraping, notice-and-consent laws against unlimited data scraping

[34] Cohen, *supra* note 12, at 1411 ("Courts treat strong data privacy protection as definitionally incompatible with constitutional speech regulation.").

[35] Bambauer, *supra* note 24, at 63. ("[f]or all practical purposes, and in every context relevant to privacy debates, data is speech").

[36] Carrero, *supra* note 10, at 151. For events taking place in public, courts also recognize a First Amendment right to record. *See* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1194-95 (2016) ("A right to record might conceivably protect a wide range of technologies for collection of public information, with the limiting case being harassment and violations of the tort of instusion on seclusion").

[37] Bambauer, *supra* note 25, at 85.

[38] Justin Marceau & Alan K. Chen, *Free Speech and Democracy in the Video Age*, 116 COLUM. L. REV. 991, 999 (2016).

may survive First Amendment scrutiny,[39] which leaves a loophole for the "right-to-record" jurisprudence. As Daniel Rauch pointed out, laws that limit data collection may curtail digital audience-information collection and potentially Digital Customized Speech.[40]

For the data creation step, the speaker's use of collected data is also affirmed as a form of speech, as they are entitled to rights to generate new ideas.[41] Data aggregation entails a right to create knowledge, which promises freedom from governments' constraints on learning something new.[42] The data creation step received more unequivocal supports than the data collection step. Some suggested that it is unlikely to enforce outright limits on use of audience information which requires to demonstrate a compelling state interests and were the least-restrictive means to do so.[43] Therefore, proposals to curtail customized speech are "neither constitutionally viable nor normatively required".[44]

However, the second viewpoint misses an important step in the assumption that data equals speech to speech equals First Amendment protection. Bambauer rationalized the existence of harmful data inferences by arguing that overprotecting "low-value and negative-value speech" will create a system where the First Amendment will impose "massive inefficiencies in our self-governance".[45] But a clear scope of overprotection for low-value and negative-value speech should be sketched. For example, the Supreme Court mentioned that speech must fit into a small number

[39] Geoffrey Xiao, *Bad Bots: Regulating the Scraping of Public Personal Information*, 34 HARV. J.L. & TECH. 701, 728-731 (2021).
[40] Rauch, supra note 1, at 438-439.
[41] Carrero, *supra* note 10, at 154-158.
[42] Bambauer, *supra* note 25, at 87-88. *See also* Ryan Calo, *Digital Market Manipulation,* 82 GEO. WASH. L. REV. 995, 1036 (2014) (arguing that Bambauer's method of seeing data is itself speech, as "distinguish[ing] gathering information from speech, while highlighting strains in First Amendment law tend in her view to bolster the case that collecting data for the purpose of speech is itself protectable speech").
[43] Rauch, *supra* note 1, at 459.
[44] *Id.* at 405.
[45] Bambauer, *supra* note 25, at 117.

of historically unprotected categories to be subject to content-based restrictions when it ruled that the Stolen Valor Act infringes on the First Amendment.[46] Free speech scholars have since overlooked digital analogies for the small number of historically unprotected categories.

We need to be aware of the capacity for the devils to be disguised under the First Amendment. It is not only low-value or negative-value speech such as fake news, commercially driven bots, or faulted algorithmic prediction that would be protected under this rationale; harmful speech that traditionally has been restricted by the First Amendment is actively seeking to revive itself under the disguise of seeing data as speech. The second viewpoint, if conducted in practice, lays a foundation for the prevalence of unbridled harmful data inferences that do not only implicate privacy harms.

iii. Third wave: limited categories of data speech enjoy free speech protection

The third-wave scholars argue that some harmful data inferences should be excluded from free speech protections and weigh *public interest* as an important indicator of content-based regulation. These scholars admit data is speech and some believe that not all forms of data as speech should be under the protection of free speech rights.[47] They may have different views on what

[46] *See* United States v. Alvarez, 567 U.S. 709, 717 (2012) ("[C]ontent-based restrictions on speech have been permitted, as a general matter, only when confined to the few 'historic and traditional categories [of expression] long familiar to the bar.'" (citing United States v. Stevens, 559 U.S. 460, 470 (2010)).

[47] Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995, 1037-38 (2014) (questioning whether harmful and manipulative marketing techniques within the context of private commercial speech should still warrant commercial platforms First Amendment protection); Richards, *supra* note 29, at 1152. ("The First Amendment critics' assumption not only ignores the reality that few data privacy rules actually involve speech, but also significantly overstates the breadth of the protection afforded by the First a protected by the First Amendment, because large categories of "speech" regulations (such as criminal solicitation, anticompetitive offers, and copyright infringement) do not in reality trigger heightened First Amendment scrutiny.")

constitutes the public interest, but they take a similar perspective toward the question of which kinds of data should be restricted; that is, to question whether data inferences are generated in accordance with the public interest.[48]

The balance between free speech protection and harmful data inferences is often cast in terms of whether such aggregated data inferences comport with the public interest. Balkin pointed out that data inferences with a higher public interest are entitled to the right to distribute information on matters of public concern, lawfully obtained, to the public.[49] From the perspective of public interest, Patel also argued that First Amendment protection should extend to civil rights testing and auditing methods to combat discriminative data inferences.[50]

First Amendment protection for for-profit data inferences is often confined.[51] Carrero proposes to draw a fine line around the scope of scraping activity with a consideration of public interest, and argues that not all forms of data scraping merit First Amendment protection, such as commercial data inferences.[52] He recommended a commercial/noncommercial classification, arguing that

[48] Mary D. Fan, *The Right to Benefit from Big Data as a Public Resource*, 96 N.Y.U. L.
REV. 1438, 1445 & 1470 (2021). ("Recognizing a right of the public to benefit … means carving out provisions for public-interest access and safe harbors that facilitate data sharing with parties qualified
and trained to protect the data and carry out research that creates public benefits such as improving health or safety. ")

[49] Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183, 1196 (2016) (citing Bartnicki v. Vopper, 532 U.S. 514, 535 (2001)).

[50] *Id.*, at 1510. ("Though this context does not necessarily need to be enforcement of civil rights statutes, it may still need to have some significant public interest element, thereby cabining the scope of this right.").

[51] Victor Brudney, *The First Amendment and Commercial Speech*, 53 B.C. L. REV. 1153 (2012).

[52] Carrero, *supra* note 10, at 166.

data scraping should be protected as a form of free speech for noncommercial actors but not commercial actors.[53] The private-sector component of data inferences, which compromises its social characteristics for the public good, may "warrant less constitutional protection" and can cause to privacy concerns.[54] In other words, it does not matter which scraping technique or how web scraping conduct is being applied; instead, it is the purpose and the context of the data scraping conduct that decide the protection of the conduct.

This article argues that the standard of public interest is a blurry and abstract concept that may not always be useful to outline the scope of protected data speech. Specifically, not only critics who rejects harmful data inferences adopted the principle of public interest, free speech advocators also appeal to the concept of public interest and the core value of free speech. For example, on the one hand, Bambauer suggests that seeing data as speech fits the longstanding vision of the First Amendment -- the creation of knowledge and the free flow of information.[55] On the other hand, Christopher Yoo also pointed out that "the editorial discretion that intermediaries exercise promotes important free speech values" through content recommendations and protecting audiences from unwanted speech such as spam, pornography, viruses, and malware.[56] If journalists and researchers are seen as harnessing data to promote accountability and democratic participation, then other data aggregators may legitimize for-profit data

[53] *Id.*, at 165-66 (In his reasoning to limit commercially oriented data inferences, Carrero contends that commercial scraping can threaten First Amendment values of intellectual privacy and should not receive the same protection as data inferences serving the public interest.).
[54] Balkin, *supra* note 51, at 1196 n. 59.
[55] Bambauer, *supra* note 25, at 63.
[56] Christopher S. Yoo, *Free Speech and the Myth of the Internet as an Unintermediated Experience*, 78 GEO. WASH. L. REV. 697, 701 (2010) (citing Eric Goldman, *Search Engine Bias and the Demise of Search Engine Utopianism*, 8 YALE J.L. & TECH. 188, 195-96 (2006); James Grimmelmann, *The Google Dilemma*, 53 N.Y.L. SCH. L. REV. 939, 941 (2008-2009)).

inferences, if they can prove they are motivated by public interest and social justice.[57]

In addition, the public interest standard does not apply to all cases, especially those commercial-driven disputes where both the plaintiff and the defendant did not show a clear inclination of public interest. Therefore, the public interest standard loses its efficiency under certain circumstances.

B. Why Free Speech Does Not Protect Certain Harmful Data Inferences

i. The problem of data anonymization

The second wave of scholarship argues that some speech that carries a risk of harm should nevertheless be protected as a necessary cost of free speech protection.[58] However, the extent of this "necessary cost" is underaddressed. The potential risks of harmful data inferences should be identified before deciding whether such risks are the necessary cost of free speech protection.

In commercial settings, as people actively seek digital resources, clandestine surveillance software collects and profiles their information into quantified datasets, which further allows digital infrastructures to sell it to third-parties. During this process, individuals become digital laborers by contributing their data to the economic growth of technology companies, even if they only clicked on recommended content and did not make any purchases through their platforms.

[57] For instance, Judge Berzon affirmed hiQ's public interests in allocating data from LinkedIn. It is also noticed that both hiQ and LinkedIn assert that "its own position would benefit the public interest by maximizing the free flow of information on the Internet.". hiQ Labs, Inc v LinkedIn Corp., 938 F 3d 985 (9th Cir 2019)

[58] Rauch, *supra* note 1, at 456 ("To be sure, empowering the marginalized carries costs; . . . . [b]ut…Protections for Digital Customized Speech serve a decisive role in letting marginalized voices be heard.").

While anonymizing data appears to be applied as a common strategy to improve data security, re-identifying a specific person with de-identified data is disturbingly easy to achieve. For example, Sweeney re-identified 87% of the U.S. population by decoding anonymized medical data with only three variables: the respondent's gender, zip code and birth date.[59] Recent research has proposed different methods and models which allow researchers to accurately estimate the likelihood of being correctly re-identified, even in a heavily incomplete dataset.[60] Even heavily sampled anonymized datasets can hardly to meet the standards of anonymized data or pseudonymous data defined in privacy laws.[61] In other words, there is a gap in the understanding of anonymized data between policymakers and data analysts. Policymakers tend to idealize anonymized data and underestimate its potential to be decrypted.[62] However, anonymized data do not hide traces of individuals; even coarse datasets provide little anonymity.[63]

While the effectiveness of data anonymization is still in question, it has become a safe harbor for technology companies. Data protection laws in the United States define aggregated data as a kind of personal information that can be used to identify specific natural persons or reflect their activities.[64] This conditional

[59] Latanya Sweeney, LABORATORY FOR INT'L DATA PRIVACY, *Working Paper LIDAP-WP4 - Uniqueness of Simple Demographics in the U.S. Population* 16 (2000).

[60] Luc Rocher et al., *Estimating the success of re-identifications in incomplete datasets using generative models.* 10 NATURE COMMC'NS, 1, 2 (2019).

[61] *See id.* at 6.

[62] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1701, 1746 (2010) (pointing out that easy re-identification will spark a frightening and unprecedented wave of privacy harm, saying, "[W]e have made a mistake, labored beneath a fundamental misunderstanding, which has assured us much less privacy than we have assumed. This mistake pervades nearly every information privacy law, regulation, and debate, yet regulators and legal scholars have paid it scant attention.").

[63] Yves-Alexandre de Montjoye, et al., *Unique in the Crowd: The privacy bounds of human mobility*. SCIENTIFIC REPS. 3(1) (2013).

[64] For instance, the Children's Online Privacy Protection Act's (15 U.S.C. §§

statement implies a critical technical uncertainty: whether data inferences can be deciphered to make a collection of non-identifiable information identifiable largely depends on the development of techniques that can be used to reverse the anonymization. In other words, if aggregated data remains anonymized and cannot be used to identify specific people, aggregated data remains outside the category of personal information and does not enjoy legal protections that apply to personal information.[65] This narrow definition of personal information, which underestimates the capacity of re-identification, affords companies a way to store personal data with less legal complication. As long as user data, such as browsing history, is aggregated or anonymized, the legal protection for such data is not as strict as personal data protection.

Technology companies seem to celebrate this safe harbor rule and adapted to it well. For example, per TikTok's privacy policy, data inferences are not under the protection of privacy policies, because legally speaking, this non-identifiable data is not seen as personal information before it is decrypted and can be used to re-identify individuals.[66] Excluding aggregated data from personal data triggers another problem: because anonymized data is not subject to privacy policies, profiling data subjects with anonymized data for behavioral advertising does not violate privacy policies. When data inferences are capable of decrypting personally

6501-06) definition of "personal information" includes a child's name, home or email address, telephone number, social security number, geolocation data, photos, videos, or audio of a child, any unique device identifier, or an IP address. The California Consumer Privacy Act (CAL. CIV. CODE § 1798.100) defines "personal information" as "information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."

[65] FEDERAL TRADE COMMISSION. *Complying with COPPA: Frequently Asked Questions*, https://www.ftc.gov/business-guidance/resources/complying-coppa-frequently-asked-questions ("COPPA only applies to personal information collected online from children, including personal information about themselves, their parents, friends, or other persons.").

[66] TIKTOK PRIVACY POLICIES, https://www.tiktok.com/legal/page/row/privacy-policy/en ("[W]e may aggregate or de-identify the information described above. Aggregated or de-identified data is not subject to this Privacy Policy.").

identifiable information, data anonymization is no longer foolproof. Therefore, current data protection policies may give users a right to anonymize personally identifiable information, but this anonymized information is substantially susceptible to the risk of re-identification.

Data inferences significantly weaken the level of protection that anonymization offers as the risk of re-identification increases. Because of the underestimated capacity of data owners' use of audience information, the second-wave scholars have generated a false impression that the created new data inference outputs directly implicate a First Amendment right, before they become aware of these anonymized data inferences might contain harms and can be reidentified. For instance, Rauch argued that states may regulate audience-information collection, but "remain almost powerless to proscribe a speaker's use of otherwise-lawfully collected audience information for political Customized Speech."[67] If legislatures do not or cannot regulate the distribution of anonymized data inferences, then users will be vulnerable to the high risk of data re-identification and privacy intrusion.

ii. Free Speech Consequentialism as a Threshold Analysis

Weighing the benefits of free speech against its potential harms has been commonly employed by courts addressing whether to preclude First Amendment protection for a form of speech.[68] Courts have applied content-based restrictions for a small number of historically unprotected categories of speech.[69] Erica Goldberg identified there is also an increasing scholarly call for the incorporation of what she termed, *free speech consequentialism* to feature the method of balancing the harms and

[67] Rauch, *supra* note 1, at 411.

[68] *See* Vincent Blasi, *Shouting "Fire!" in a Theater and Vilifying Corn Dealers*, 39 CAP. U. L. REV. 535, 537-48 (2011); Rebecca L. Brown, *The Harm Principle and Free Speech*, 89 S. CAL. L. REV. 953, 1003 (2016).

[69] *Alvarez*, 567 U.S. at 717.

benefits of speech.[70]

Free speech consequentialism provides a valuable account for assessing the scope of unprotected data speech outside of the First Amendment protection. On this view, scholars embraced two considerations. First, they believe a balancing inquiry serves as means to achieve free speech's ends, whether for truth or democratic self-government.[71] This harm-based balancing approach entails a judgement that at least some types of harmful speech do not implicate free speech rights. Having recognized this point, the next step is to decide the extent of harms to be assessed as harmful enough to preclude First Amendment protection. Certain scholars seek to limit the legally recognized harms of free speech to conduct harms, or harms that "elicit physical or tangible responses or impairment of immediate, material interests."[72] This narrow category does not include incendiary effects, broken social norms, or grave emotional harms without tangible interests.[73]

Discussions on data inferences and free speech have repeated this balancing approach offered by free speech consequentialism. For instance, Rauch argued that fears about citizen autonomy and hyper- partisan factions caused by political speech customization can be offset by associated benefits of empowering marginalized communities and checking government power.[74] The harm-based approach to data inferences, as a variant of free speech consequentialism, reflects a premise that the protection of harmful speech is an inevitable feature of free speech rights.

[70] Goldberg, *supra* note 21, at 689.

[71] *See id.* at 695-701; Cohen, *supra* note 12, at 1410-1411(Citing *Central Hudson*, which summarized four-part test for the balancing inquiry: "if the regulation targets a communication that is not misleading or related to unlawful activity, it must be supported by a substantial government interest, must materially advance that interest, and must not be more restrictive than necessary to serve that interest.").

[72] Goldberg, *supra* note 21, at 730.

[73] Massaro, Norton & Kaminski, *supra* note 20, at 2501 ("First Amendment protections are applicable to racist, homophobic, sexist, blasphemous or otherwise cruel postings on Facebook or other social media sites.").

[74] Rauch, *supra* note 1, at 413.

Nevertheless, protecting the good speech does not need to be upheld at the expense of including valueless data speech. This article argues that the beneficial uses of customized speech should not be seen as "offsets" because this view implies that in order to preserve the benefits of customized speech, we should keep the harmful ones. As Leslie Kendrick put it, strong protections for free speech need not "begin with the premise that free speech must protect harmful content in order to be meaningful."[75]

A vital difference between Rauch's application of the harm-based approach and free speech consequentialism is that Rauch made a judgement about the whole category of Customized Speech by raising the single category of political speech customization and analyzing its harms and benefits.[76] Compared with Rauch's approach, Goldberg's free speech consequentialism applies on a case-by-case basis. The inductive reasoning of the former becomes especially dangerous when it meets a mixed-purpose incident in which data inferences are not only political data speech, but also devil data speech with other malicious features. In other words, drawing a conclusion about a particular category of speech should also consider its relevant characteristics: if one characteristic of the data speech goes beyond legality, then it does not matter how many other legitimate and beautifully-envisioned characteristics it embodies — it should not be considered free speech. In the case of Cambridge Analytica, free speech welcomes political speech, but this does not exempt scrutiny from a commercial data speech perspective, especially when it involves manipulative marketing techniques.

Free speech consequentialism is not a perfect theory to determine a precise rule for how free speech should be applied to data inferences, although Goldberg argues that courts often apply

[75] Kendrick, *supra* note 16, at 105.

[76] Rauch, *supra* note 1, at 405, 407 (defining Customized Speech as "speech targeted or tailored based on knowledge of one's audience" but concluding that "the use of audience information to customize speech is, itself, core protected speech" solely based on discussions on political customized speech.)

this approach. [77] But it is an ideal theory for identifying a threshold for malignant speech. Nevertheless, as pointed out by Amanda Shanor, a deeper examination of the sociological understanding of speech that goes beyond the calculation of harms and benefits may explain when free speech consequentialism gets carried out and when it does not.[78] This paper will elaborate a listener-centric approach to compensate the sociological influences of free speech consequentialism in Section IV.

III. A SPECTRUM

Although we accept the proposition that generalized data should be seen as speech, this does not mean that data inferences cannot create social bias or invade privacy simply because they are a form of speech. The question of whether data inferences fall outside the bonds of First Amendment protection should be subsumed under a larger set of circumstances, such as commercial data inferences and discriminative data inferences.

A. CLASSIFICATION OF DATA INFERENCES

This section posits a spectrum theory that views data inferences as occurring along a continuum of actions. On the one end of the spectrum are the most protected data inferences, including individual rather than commercial works. They may contribute to the public interest, like journalistic or scholarly research, but those individual works that are deemed valueless and harmful data speech also belong to this category. These data inferences are fully covered by the First Amendment. The data inferences situated on the other end of the spectrum represent harmful, malicious data speech inferences that would cause imminent, tangible harm that bear resemblance to conduct harm. These data inferences that

[77] Goldberg, *supra* note 21 at 687 ("Under current doctrine, courts determine if speech can be regulated using various forms of free speech consequentialism, such as weighing whether a particular kind of speech causes harms that outweigh its benefits, or asking whether the government has especially strong reasons for regulating particular kinds of speech.")

[78] Amanda Shanor, *First Amendment Coverage*, 93 N.Y.U. L. REV. 318, 343-346 (2018).

carry the consequences of conduct harm should not be protected by the First Amendment. Althoughdata inferences along the two extreme ends are easily identifiable, most data inferences occur in the middle between the two ends, predominantly in commercial speech. This section classifies four types of data speech: (1) individual data speech and (2) devil data speech at two ends of the spectrum, and then (3) commercial data speech and (4) political data speech, which represent the center categories.

i. Devil Data Speech

The historical doctrine of free speech consequentialism goes back to Justice Holmes' classic statement of the clear and present danger test: "The question in every case is whether the words used are used in such circumstances and are of such a nature as to create a clear and present danger that they will bring about the substantive evils that Congress has a right to prevent."[79] Based on Holmes's characterization of harmful speech, this paper names the kind of data speech that creates an imminent, tangible danger that will bring about enough "substantive evils" to be precluded outside of First Amendment protection devil data speech.

Using the example of deepfake videos, this section argues that legislators and policymakers should not conclude whether data inferences are under free speech protections without regard for the characteristics and purposes of data inferences. That being said, the characteristics of the different categories of data inferences under the First Amendment should be defined under specific contexts. For example, data speech like deep fake nonconsensual pornography that causes an imminent and tangible danger should not enjoy First Amendment protection.

Over 90% of deepfake video technology application is used to create nonconsensual fake pornography.[80] Harmful data

[79] Schenck v. United States, 249 U.S. 47, 52 (1919).
[80] Karen Hao, *Deepfake porn is ruining women's lives. Now the law may finally ban it*, MIT TECH. REV., (February 12, 2021), https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

inferences generated through deepfakes are a classic example that runs counter to the First Amendment's goal of protecting truth-seeking.[81] These data inferences, after all, have nothing to do with finding scientific, sociological, or political truth. The ability to generate factually distorted content by synthesizing existing data usually in the form of video presentations has stoked modern society, mostly in a bad way. Of course, there are positive sides to this tech innovation. For instance, the Dutch police generated deepfake videos to help draw attention to a murder case, hoping to attract witnesses.[82] But just as people cannot be forgiven of felonies because they used volunteer in the community, deepfakes cannot get rid of their sins through rarely occurring benevolent applications.

Scholars have appealed to regulate the harmful content deepfakes generate. For instance, Jared Schroeder argued that deepfakes should be subject to limitation as they pose a threat to democratic discourse.[83] Franks and Waldman added that deepfake manipulation is a form of deliberatively deceptive speech that is disproportionately designed to target women and the queer community, rendering it especially dangerous to vulnerable social groups.[84] And even if no sexual violence is involved, some types of deceptive images can be exploited to "threaten, intimidate, and

[81] The truth theory of free speech puts forward a "marketplace of ideas" theory. Although it can be traced to political philosophy of John Stuart Mill in his *On Liberty*, this theory was first raised in law by Justice Holmes in in *Abrams v. United States*: "When men have realized that time has upset many fighting faiths, they may come to believe even more than they believe the very foundations of their own conduct that the ultimate good desired is better reached by free trade in ideas – that the best test of truth is the power of the thought to get itself accepted in the competition of the market." 250 U.S. 616 (1919).

[82] Lara Smit, *Dutch police use deepfake technology in bid to solve 19-year-old cold case murder of Sedar Soares*, ABC NEWS (May 26, 2022), https://www.abc.net.au/news/2022-05-26/murdered-teen-asks-public-for-help-in-deepfake-video/101100874.

[83] Jared Schroeder, *Free Expression Rationales and the Problem of Deepfakes within the E.U. and U.S. Legal Systems*, 70 SYRACUSE L. REV. 1171 (2020).

[84] Franks & Waldman, *supra* note 6, at 894.

inflict psychological harm on the individual depicted."[85] Even free speech scholars cannot deny that these malicious data inferences have low social value and exacerbate truth decay.[86]

A critical consideration for excluding deepfaked pornography from of free speech protections is that the harms they carry go beyond privacy issues and cause real threats to people's lives. *Planned Parenthood v. American. Coalition of Life Activists* serves as an example. The American Coalition of Life Activists posted the names and locations of abortion providers to a website, which led to the murder of three abortion providers. The posters' content was not protected by the First Amendment because it constituted a "true threat."[87] Similarly, it is hard to believe that defamatory content in a deepfaked video would not cause reasonable fear of serious bodily injury or substantial emotional distress that results in real threats to people's lives. To substantiate the legitimacy to criminalize the course of cyber misbehavior that would cause fear, Chesney and Citron raised the federal cyberstalking statute, 18 U.S.C. §2261A, which highlights that it is a felony to use "interactive computer service or electronic communication service" to "intimidate" victims or their immediate family in ways "would be reasonably expected to cause substantial emotional distress."[88]

Deepfaked data inferences is not the only category of harmful data inferences. The tangible harms caused by data inferences with discrimination in terms of race, gender and ethnicity is a fact that is constantly reminded by scholars and journalists since

[85] Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1774. (2019).
[86] *See* Robert Chesney & Danielle Citron, *21st Century-Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security*, 78 MD. L. REV. 882, 885-88 (2019), for a survey of distrubing deepfakes.
[87] Planned Parenthood of Columbia/Willamette, Inc. v. Am. Coalition of Life Activists, 290 F.3d 1058, 1086 (9th Cir. 2002).
[88] Chesney & Citron, *supra* note 86, at 1801-1802.

the popularization of algorithms.[89] Free speech enthusiasts for data inferences, particularly scholars supporting the second viewpoint we discussed above, need to address the problem raised in such scenarios: how can free speech justify its protections to data speech with biased results which cause harmful real-life consequences?

Under the model of seeing the degrees of free speech protections for data inferences as a spectrum, devil data inferences like deepfaked pornography is situated upon the absolute unprotected end. This proves that, at least, there are some kinds of data inferences in the form of video expression that should not be protected by the First Amendment. To say that protected speech does not have to be fact-driven or merit-based to be covered by the First Amendment does not necessarily lead to the conclusion that harmful data speech can exploit the value of data and wreak havoc on people's lives under the shield of free speech rights. The bottom line of data speech is whether it causes unacceptable imminent dangers to people's safety and livelihoods, including but not limited to death threats, loss of education and job opportunities as a result of algorithmic bias, and deepfaked videos that irreversibly tarnish one's reputation.

ii. Individual Data Speech

For the types of data speech situated on the protected category of the spectrum, we propose that it may not always cater to public interest, but it should be generated by individuals rather

[89] The countless cases in which machine learning techniques discriminate against disadvantaged social groups in terms of gender and race have proved that data inferences have the capability to implement people's hidden biases in algorithms and results. For a critical view of how shared ownership of identify is ignore in data protection laws, *see* Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30 PHILO. TECHNOL. 475, 479 (2017), in which Mittelstadt notes that "patterns and correlations used to group individuals are functionally equivalent to identifiers but are not afforded comparable status under existing data protection law."

than commercial entities.[90] This consideration is grounded in the idea that in order to align with the fundamental human right of free speech, data owners must take data subjects' human dignity into consideration, which has long been recognized as resisting incorporation into commercial entities or market efficiency terms.[91]

Unlike deepfakes, data inferences for the public interest conducted by individuals have been embraced among scholars. If the degrees of free speech protection received by data inferences can be characterized as a spectrum with deepfaked pornography at one end that avoids free speech protection, then the other end would be scraping and generating data inferences for the public interest, usually conducted by journalists and researchers at public institutions. For example, although both employed data scraping techniques to analyze people's behavior on Facebook, researcher Jonathan Albright's scholarly scraping behavior is widely considered consistent with the First Amendment values of democratic self-governance and autonomy,[92] whereas Cambridge Analytica's right to receive free speech protections[93] was severely challenged. Both courses conduct scraped, and aggregated information obtained on Facebook, but they are attached with different significance. In scholarship, data inference conduct serving the public interest is more likely to be considered protected under free speech doctrine.

While it has long been a fixture that actions concerning the public good with nonprofit motives are more likely to be included in the free speech considerations, we try to set the lower limit of the protected category of data inferences. It is argued that false data speech with no explicit public interests coming from individuals also warrants free speech protection on a human dignity

[90] Not all individual data speeches are on the end of absolutely protected data speech, but the absolutely protected data speech on the spectrum should represent individual freedom.
[91] Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1260 (1998).
[92] Carrero, *supra* note 10, at 132-133.
[93] Rauch, *supra* note 1, at 409.

ground.

Freedom of expression is given as a fundamental human right in various international legal documents, such as the UN Declaration of Human Rights, the International Covenant on Civil and Political Rights, and many other international and regional treaties. It underpins human dignity and other key values, such as freedom of association and freedom of the press.[94] It is, along with privacy, a human-dignity-driven right serving the interests of the public rather than the interests of private conglomerates. At the theoretical level, Edwin Baker sees individual self-fulfillment and participation in change as the key First Amendment values.[95] Steven Heyman also pointed out that the basis of individual rights is founded on human dignity and inherent freedom in the natural rights tradition.[96]

We find human dignity's incarnation in American case law.[97] The reinforcement of self-actualization and individualistic perspectives of free speech can be found in the Supreme Court's jurisprudence. In *Mosley*, Justice Marshall suggested that granting people with freedom of expression is "to issue self-fulfillment for each individual." [98] Justice Harlan's characterized the First

[94] The Universal Declaration of Human Rights opens with an acknowledgement of "[w]hereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world." Universal Declaration of Human Rights, G.A. Res. 217A(III), at 71, U.N. GAOR, 3d Sess., 1st plen. mtg., U.N. Doc. A/810 (Dec. 12, 1948).

[95] Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964,
991 (1978).

[96] STEVEN HEYMAN, FREE SPEECH AND HUMAN DIGNITY (2008), at 38.

[97] Mich le Finck, *The Role of Human Dignity in Gay Rights Adjudication and Legislation: A Comparative Perspective*, 14 INT'L. J. CONST. L. 26, 32-39 (2016) (discussing the resort to human dignity in US gay rights litigations); David C. Yamada, *Human Dignity and American Employment Law*, 43 U. RICH. L. REV. 523(2009) (discussing dignity claims in US employment rights).

[98] HEYMAN, *supra* note 101, at 82.

Amendment as designed to "comport with the premise of individual dignity and choice upon which our political system rests."[99] An individual's autonomy is impaired when the government restricts speech because it disapproved of an individual's decision to express them. However, this situation does not work for a commercial entity's speech because data inferences are not reflections of a company's dignity and belief.

Such individual-based human dignity is mutually exclusive to commercial-based speech. Although data inferences may reflect companies 'dignity and belief, these commercial entities are not moral agents and reflect no human dignity. An assumption behind the expression of human dignity is that it is not possessed by private companies.[100] Although a corporation can enjoy the freedom of speech,[101] it should not be included as an agent upholding human dignity. Moreover, we must note that data generated by private companies can sometimes even pose threats to human dignity. For instance, during the observation stage of information collection, nonconsensual and extensive information collection runs contrary to human dignity.[102]

iii. Commercial Data Speech

Commercial data speech is the broadest and most complicated data speech category. Once speech becomes commercialized, its purposes and goals with attached financial motives may delegitimize its entitlement to free speech protections. Jack Balkin warns that just as companies would employ the First Amendment to advance their free speech rights, they would also seek constitutional protection for surveillance and social control to promote

[99] *Id.* at 85 (citing Cohen v. California, 403 U.S. 15 (1971)).

[100] Luciano Floridi, *On Human Dignity as a Foundation for the Right to Privacy.* 29 PHIL. & TECH., 307-312 (2016).

[101] Citizens United v. Fed. Elec. Comm'n, 558 U.S. 310 (2010) (holding that the First Amendment protects independent expenditures for political campaigns by corporations as speech).

[102] Kang, *supra* note 91, at 1260.

business models and protect their profits.[103]

Although commercial speech is not situated on the end of unprotected speech, courts have historically afforded commercial speech a limited degree of protection due to "its subordinate position in the scale of First Amendment values."[104] With a content-based approach, courts usually assess if the content is made for profit reasons,[105] or if the content is neutral,[106] to measure whether commercial speech is an appropriate fit for free speech protections.

Victor Brudney offers a vivid comparison of two cases, *In re Primus*, 436 U.S. 412 (1978) and *Ohralik v. Ohio State Bar Ass'n*, 436 U.S. 447 (1978) to illustrate how commercial speech with a financially driven incentives has compromised the legitimacy of free speech.[107] In *Primus*, the Court found that South Carolina's application of disciplinary action to an attorney affiliated with the American Civil Liberties Union, charging the attorney with solicitation for helping a sterilized woman sue her doctor, violated the First Amendment.[108] Yet in a commercial for-profit setting, solicitation by another attorney, Ohralik, was not protected by the First Amendment despite no finding of actual harm because his solicitation "was not engaged in associational activity for the advancement of beliefs and ideas; his purpose was the advancement of his own commercial interests."[109]

Therefore, even commercial speech with an informational and educational function cannot avoid the question of whether it is

[103] Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation,* 51 U.C. DAVIS L. REV. 1149, 1187 (2018).
[104] Ohralik v. Ohio State Bar Ass'n, 436 U.S. 447, 456 (1978).
[105] See the discussion of *In re* Primus, 436 U.S. 412 (1978) and Ohralik v. Ohio State Bar Ass'n, 436 U.S. 447 (1978) *infra* at 18.
[106] Daniel A. Farber, *Commercial Speech and First Amendment Theory*, 74 NW. U. L. REV. 372, 374 (1979-1980).
[107] Victor Brudney, *The First Amendment and Commercial Speech*, 53 B.C. L. REV. 1153, 1191-1199 (2012).
[108] In re Primus, 436 U.S. 412, 431 (1978).
[109] Ohralik., 436 U.S. at 458-59.

protected by the First Amendment.[110] And for data inference speech that does not benefit the public interest except for participating in the data economy, it is necessary to examine the contexts where data cannot be recognized as free speech. This article will elaborate on the conditions that limit the application of commercial data inferences in Section IV.

iv. Political Data Speech

The most controversial example of political data speech so far is probably the scandalous Cambridge Analytica incident. Cambridge Analytica showcases that despite their political nature, many instances of political data speech also embody other characteristics, such as commercial data speech. For the political data inference that is also commercial data speech, the assessment of whether such political data speech qualifies for free speech protection needs to take into account the stealth and even unlawful application of forbidden marketing techniques, like deceptive advertising and manipulative speeches.

With permissions granted by Facebook, Cambridge Analytica was able to access and collect personal information from 71.6 million Facebook users. It is recognized that Cambridge Analytica was able to predict and influence choices at the ballot box through such a large dataset. Many argue that such expression should be exempted from First Amendment coverage on the ground that it appears detrimental to democratic society,[111] while

[110] Zauderer v. Off. of Disciplinary Counsel of the Sup. Ct. of Ohio, 471 U.S. 626, 651 (1985) ("[T]he extension of First Amendment protection to commercial speech is justified principally by the value to consumers of the information such speech provides."); Va. State Bd. of Pharmacy v. Va. Citizen Consumer Council, Inc., 425 U.S. 748, 763-64 (1976) (emphasizing the value of "the free flow of commercial information" to individual consumers and the public more generally).

[111] JARON LANIER, TEN ARGUMENTS FOR DELETING YOUR SOCIAL MEDIA ACCOUNTS RIGHT NOW 110 (2018); Christopher S. Elmendorf & Abby K. Wood, *Elite Political Ignorance: Law, Data, and the Representation of (Mis)perceived Electorates*, 52 U.C. DAVIS L. REV. 571, 607 (2018); Gregory P. Magarian, *How Cheap Speech Underserves and Overheats Democracy*, 54 U.C. DAVIS L. REV. 2455, 2470 (2021).

Rauch [112] pointed out that in order to preserve the virtue of free speech, the harms Cambridge Analytica caused can be offset by benefits of seeing political data speech under free speech protections.

Legal tolerance to incendiary or dangerous political speech, although not necessarily to the level of immunity, reflects a higher degree than similarly risky non-speech activity.[113] Shanor suggested that the First Amendment extends its coverage to political speech because when there is a pluralistic interpretive community, the First Amendment generally offers its coverage.[114] Furthermore, different standards of assessment of the effect of the political speech activity at issue are the result of a pluralistic interpretive community.[115]

However, both speakers and listeners face a much more complicated situation in the context of data speech. For example, fake news can be seen as "individuals seeking simultaneously to distinguish themselves through individualization or self-identification and to connect themselves through group association with a community of people" based on shared values.[116] Nonetheless, with bad commercial bots taking up a quarter of internet traffic, it seems difficult for fake news to help individuals to achieve self-identification. [117] When people think they are bonding with a community of people, in fact they are only interacting with a community of robots.

[112] Rauch, *supra* note 1, at 413.
[113] Kendrick, *supra* note 15, at 106.
[114] Shanor, *supra* note 79, at 354.
[115] *Id.* at 352.
[116] Alan K. Chen, *Free Speech, Rational Deliberation, and Some Truths about Lies*, 62 WM. & MARY L. REV. 357 (2020) ("The regulation of political data speech could be conceptualized in a manner not that different from censoring art, or video games all of which might create or inspire a different or alternative worldview. Allowing people to alter their opinions at their free will have their social values in the sense that it enhances individuals' ability and freedom to internally experience self-realization.").
[117] Erez Hasson, *Bad Bot Report 2021: The Pandemic of the Internet*, IMPERVA (April 13, 2021), https://www.imperva.com/blog/bad-bot-report-2021-the-pandemic-of-the-internet/.

Political data speech should be carefully analyzed, especially those that are also commercial data inferences, before being treated as general political speech. Political data speech generated by individuals is more likely to be consistent with the core value of free speech than its commercialized counterparts. For political data speech with a mixed purpose like the Cambridge Analytica incident, its eligibility to be included in free speech protection should be examined *after* it is confirmed as lawful under commercial laws. For example, in examining whether free speech rights can be applicable to the Cambridge Analytica case, the factor of its marketing tactics should be taken into account. Cambridge Analytica's fitness for First Amendment coverage should be questioned, not because its attempted goal favored a president that might "hurt democracy," but because the voter manipulation it engaged.[118] The data collection Cambridge Analytica applied is a form of manipulative speech in the category of commercial speech, exploiting Facebook users' knowledge vulnerabilities. During this process, listeners' interests were compromised as speakers covertly influenced those listeners' choices to the speakers' advantage without the listeners' conscious awareness.[119] This covert manipulation of people's judgement delegitimized the data speech's free speech rights.

The above examples of civil rights tests, deepfakes videos, and discriminatory data inferences illustrate that different forms of data speech require different levels of regulatory enforcement. None of these data inferences can solely define how data inferences should be protected by the First Amendment. The answer to that question depends on the type of data inference and its location

[118] Caitlin Dewey, *Facebook fake-news writer: 'I think Donald Trump is in the White House because of me'*, WASH. POST (Nov.17, 2016), https://www.washingtonpost.com/news/the- intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white- house-because-of-me/; also see Daniel Susser, Beate Roessler & Helen Nissenbaum, *Online Manipulation: Hidden Influences in a Digital World*, 4 GEO. L. TECH. REV. 1, 28-29 (2019).

[119] Jones, *supra* note 24, at 510.

along the spectrum. However, when scholars reach conclusions regarding the relationship of free speech and data inferences, the contexts they apply are different. For example, when Rauch argues that digital customized speech promotes democratic values, he mainly refers to political data speech instead of other, worse data inferences, such as deepfake technologies. Or, when Franks and Waldman warned us that the unbridled use for free speech with data inferences has led to "the disintegration of truth, the reign of unanswerable speech, and the silencing and self-censorship" of social minorities,[120] they depended solely on the case of deepfaked videos to reach this conclusion.

It seems farfetched to rely on the example of sock puppet speech to give full support to the argument that data inferences equal all free speech without considering other types of data inferences, such as discriminatory speech, hate speech, or conduct-alike harmful speech. It is also not reasonable to select a few cases such as political speech or good-intentioned deepfakes to argue that all data inferences should be protected as free speech. This inconsistency in defining data inferences can lead to different conclusions. It also reminds us that a classification of data inferences is a precondition to understanding the relationship between data inferences and free speech. A classification of data inferences must be done before considering whether they merit First Amendment protection. All in all, data inferences can be seen as a form of speech, but certain forms of data speech should be restrained. The next section dives into the category of commercial speech and analyzes how it should be evaluated under the First Amendment, with an emphasis on listeners' interests.

IV. A NEW PROPOSAL: REGAINING DIGITAL RIGHTS

As discussed above, profit-driven commercial speech faces more stringent standards for free speech protections than noncommercial speech. This section addresses the data inferences in the middle of the spectrum, which require further scrutiny, and specifies the circumstances under which their free speech claims fail or

[120] Franks & Waldman, *supra* note 6, at 896.

succeed.

To do so, we propose three key questions regarding the ethical application of data generalization. Do commercial data inferences encompass the intrinsic public interest? Do commercial data aggregations bring with them a result of discrimination or social injustice? Do digital users provide consent to making data inferences?

A. THE LENS OF LISTENERS: A LISTENER-CENTRIC APPROACH

It is widely recognized that listeners doctrinally enjoy no less freedom of speech than speakers.[121] Doctrinally speaking, listeners' interests are mirrors of the social effects of speech. When Shanor argued for a deeper examination of the sociological influence of speech that goes beyond harms and benefits, she expressly noted that the pattern of First Amendment coverage can be explained by a sort of social consequentialism and a kind of "speech effect," including how a speaker can affect the behavior of or cause harm to a listener and how the activity in question influences its surrounding social dynamics and social norms.[122] Han also included the consideration of social harm on audiences, arguing that courts should measure foreseeable social harm that would likely be elicited through audiences' processing of a particular type of speech.[123]

In fact, courts have interpreted the First Amendment to permit the government to "intervene on listeners' behalf by prohibiting false and misleading speech, requiring speakers to stay away

[121] James Grimmelmann, *Listeners' Choices*, 90 U. COLO. L. REV. 365, 401 (2019); C. Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964, 1007 (1978) ("[T]he listener has a right to demand that the government not prohibit the listener from receiving or using information."); Helen Norton, *Truth and Lies in the Workplace: Employer Speech and the First Amendment,* 101 MINN. L. REV. 31, 55 (2016).
[122] Shanor, *supra* note 79, at 344-345.
[123] David S. Han, *The Mechanics of First Amendment Audience Analysis*, 55 WM. & MARY L. REV 1647, 1682-83 (2014).

from listeners who prefer to be left alone or requiring speakers to make accurate disclosures of material matters."[124] The influence on listeners sometimes plays a decisive role in measuring the level of scrutiny for the speaker. For example, in *Zauderer v. Office of Disciplinary Counsel*, the Supreme Court held that commercial disclosure requirements served consumers' information interests as listeners.[125] The example of the advertising law is also telling: commercial actors are mandated to provide accurate disclosures about their products. This suggests that to cope with harmful influences of some types of information, the government is allowed to intervene to protect vulnerable listeners under First Amendment law.[126]

However, listeners' interests are often less emphasized in the context of data speech. In the current climate wherein users' free speech rights are undervalued, massive commercial data leaks are relatively common. In a secret deal, Ascension, one of the largest healthcare providers in the U.S, transferred the medical data of 50 million Americans to Google. A whistleblower involved in the deal reported that "[p]atients haven't been told how Ascension is using their data and have not consented to their data being transferred to the cloud or being used by Google."[127] Commercial data inference cases such as the Cambridge Analytica and *hiQ* have shown that data speakers who aggregate data without listeners' conscious awareness frustrate listeners interests. The role of the user is neglected despite the fact that "members own the content and information they submit or post to LinkedIn," and LinkedIn only functions as an intermediary to use and distribute that information.[128] Unlike other types of speech, the uniqueness of data inferences as a form of speech lies in the power asymmetry between

[124] Norton, *supra* note 29, at 231.
[125] 471 U.S. 626, 651-52 (1985).
[126] Norton, *supra* note 105, at 232-237.
[127] Ed Pilkington, *Google's Secret Cache of Medical Data Includes Names and Full Details of Millions – Whistleblower*, GUARDIAN (Nov. 12, 2019), https://www.theguardian.com/technology/2019/nov/12/google-medical-data-project-nightingale-secret-transfer-us-health-information.
[128] hiQ Labs v. LinkedIn, 31. F.4th 1180, 1185 (9th Cir. 2022) (citing LinkedIn's User Agreement).

the data controllers who generate and analyze data speech and the data subjects who are likely to be unaware of their data analysis results.

As mentioned in Section II, data protection laws do not consider non-identifying information as personal information. If we read through the privacy policies of mainstream apps carefully, we would find that most user-generated data, like browsing and search histories, are usually anonymized to qualify as non-identifiable data and are thus not subject to privacy policies, at least before they are decrypted to be identifiable. For aggregated data, there is no legal obligation for platforms to inform the users that their data is being processed and handled to a third party because their data are already anonymized.[129] This creates a huge information power asymmetry between the data subjects and the data processors. Users' interests are greatly compromised.

That asymmetry is intensified given that free speech protections shield harmful data inferences on the platform. The data subjects' freedom of speech, in turn, is weakened by the platforms' chilling effect under the digital panopticon. Indeed, modern privacy problems can create self-restraint and self-censorship.[130] Although governmental surveillance produces the greatest chilling effect, non-state actors can also monitor people's legal online activities.[131] When the willingness of listeners to disclose information decreases due to fear of data collection and distribution, it means the increase in data collectors' freedom of speech trumps the free speech values that protect data subjects' interests.

Therefore, we argue for reviving a listener-centered principle for assessing whether a particular category of commercial

[129] Even the European Union's General Data Protection Regulation – among the most stringent data protection rules – states that it does not "concern the processing of such anonymous information," and that the principles of data protection should not apply to anonymous information. Council Regulation 2016/679, 2016 O.J. (L 119) 1, 5 (Recital 26).

[130] DANIEL SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 33-35 (2006).

[131] Penney, *supra* note 17, at 5.

data speech should be included under the First Amendment's absolute protections. With a listener-centric principle, commercial data inferences should realign themselves through three perspectives: (1) whether they contribute to the public interest; (2) whether they can be counter-argued; (3) whether they acquire consent from the users.

i. Who Benefits From Data Generalization?

While some have argued that data scraping that serves the public interest conducted by researchers and journalists merits First Amendment protection, we cannot preclude the possibility that commercial data inferences may contribute to the public interest, especially given the fact that "tech for good" projects continue to mount. It is important to remember, however, that manipulative commercial speech which is generated by frustrating listeners' interests should be treated as entirely unprotected by the First Amendment.

When we evaluate whether free speech is applicable to commercial speech, the social contribution and public interest that the commercial data inference speech carries play an important role. For example, some apps donate a certain amount of money to charity based on the accumulated steps people walk with a wearable device or a smartphone.[132] Tracking users' steps allow these apps to monitor walking patterns to estimate users' health. This data can be valuable when aggregated to evaluate a particular health indicator of people living in a certain area. These apps with records of people's physical activities can even allow one to raise money for charity.[133]

Giving commercial data inferences that bear obvious public interest a higher priority does not mean commercial speech without public interest falls outside the First Amendment's reach.

[132] Joe Lepper, *The Best Tech for Good Projects Happening now*, CHARITY DIGITAL (July 8, 2020), https://charitydigital.org.uk/topics/topics/the-best-tech-for-good-projects-happening-now-6525.

[133] *See, e.g.* the app Charity Miles, https://charitymiles.org/.

As Frederick Schauer observed, the question of which forms of speech can be covered by the First Amendment is a separate question from the question of how much protection such speech should receive.[134] Narrow commercial speech which does not serve anyone's autonomy interest can still receive free speech coverage; but the level of protections they receive is of a lower degree than speech that enhances listeners' autonomy.

ii. Whether Harmful Data Speech can be Counter-argued

One of the most commonly used arguments to support low-value or even harmful data inferences is the theory of the "marketplace of ideas" and the belief that an unfettered marketplace of ideas ultimately leads to the discovery of truth.[135] Accordingly, the regulation of data speech is assumed to be the main threat to the marketplace of ideas, and precluding things like prior restraints on on publication.[136] However, the market is different in the case of data inferences, as data inferences that comprise about 25% of website traffic reflects mostly commercial incentives.[137] If the idea of the marketplace allows these commercial incentives to grow unfettered, they will drown out human voices. The essence of the marketplace of ideas, in this case, can only be achieved if these commercial bots are carefully evaluated.

The concept of the marketplace of ideas is a contextual one. Put forward a century ago, it mainly applies to the context in which everyone is free to voice their opinions, not just the richest advocator who can dominate the market with their opinions using capital force. But now, in the era of scraping bots and commercial data

[134] Frederick Schauer, *Out of Range: On Patently Uncovered Speech*, 128 HARV. L. REV. F. 346, 348 (2015).
[135] Michael Parsons, *Fighting for Attention: Democracy, Free Speech, and the Marketplace of Ideas*, 104 MINN. L. REV. 2157, 2238–39 (2020) (discussing the potential value for the marketplace of ideas that legislators can extrapolate through regulation of private company data use in advertising).
[136] Wu, *supra* note 23, at 554 (pointing out an assumption that the marketplace of ideas would operate well by itself without government intervention).
[137] Hasson, *supra* note 124.

inferences, the data subjects who avoid disclosing more information remain powerless in front of digital platforms' speech aggregation outputs, after they signed the user agreement which does not protect aggregated data the same way as it protects personal data. As the Supreme Court has held, "false statements of fact are particularly valueless; they interfere with the truth-seeking of the marketplace of ideas, and they cause damage to an individual's reputation that cannot easily be repaired by counter-speech, however persuasive or effective."[138]

Another argument often employed to support the marketplace of ideas is that harmful speech is always best addressed through counter-speech rather than regulation. However, commercial speech itself is less likely to be confronted by counter or corrective speech compared to other types of speech protected under the First Amendment.[139] In the context of harmful data inferences, not only is it difficult to correct data inferences, but many are essentially undetectable.[140] Harmful data inferences about individuals such as deepfaked pornography may not be noticed by data subjects and therefore it is less likely the victims will address these fraudulent representations with equivalent speech. Like the unauthorized publication of a victim's name, the dissemination of a home address, or the disclosure of one's sexual orientation, data inferences are not ideas that can simply be countered with different and better ideas. Harmful data inferences from speech that cannot be counterargued should be considered beyond the scope of First Amendment coverage.

iii. Whether Data Subjects Authorize Consent

To avoid the misuse of data inferences, scholars have called for platforms to leverage more accountabilities. But at the same time, the task to diagnose risks for harmful data inferences is

[138] Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52 (1988).

[139] Victor Brudney, *The First Amendment and Commercial Speech*, 53 B.C. L. REV. 1153, 1154 (2012).

[140] Franks and Waldman, *supra* note 6, at 895.

delegated to platforms, instead of having users themselves in control. For instance, Wachter and Mittelstadt propose assigning the data controller a role of evaluating whether such inferences are reasonable.[141] Balkin also suggested considering digital infrastructures as information fiduciaries that share the duty of safeguarding the sensitive information of end users.[142] The fiduciary role of social media service providers exemplifies how commercial data can be motivated to align with the public interest by restricting the infrastructures' capacity to collect and analyze personal information.

However, the impact of information fiduciaries is limited in the case of data generalization because of the innate conflict between commercial interests and clients' privacy. Balkin pointed out the power asymmetries between the fiduciary who collects and operates upon sensitive information about the client, which leaves clients in a position where they have to trust fiduciaries and "hope that the latter will not betray them".[143] He raised professional such as doctors and lawyers as the classic examples of information fiduciaries.[144]

Unlike other circumstances that Balkin raised wherein the clients' interest aligns with the service provider's interest — helping clients to protecting privacy is establishing authorities and making more money for the fiduciaries like doctors and lawyers themselves, generating data inferences is a zero-sum process in which service of the clients' interest, from the very least of privacy to a potential risk of algorithmic discrimination, decreases with the increase in disclosed information from data inferences. The critical difference between information service providers and other professionals is that the former distinctively rely on a business of extracting information provided by clients. Information fiduciaries cannot be optimized because technology companies will always

[141] Wachter & Mittelstadt, *supra* note 2, at 613. (proposing a new right to reasonable inferences applicable to high risk inferences that cause damage to privacy or reputation, which "would require *ex-ante* justification to be given by the data controller to establish whether an inference is reasonable").
[142] Balkin, *supra* note 36, at 1186.
[143] *Id.* at 1160.
[144] *Id.* at 1161.

operate under the logic of prioritizing their commercial interests, not the clients' interests.

Having data controllers assess the adequacy of data inferences is like having someone be the judge and game player at the same time. Data controllers are always inclined to explain the usage of data inferences to their advantage. Facebook has shown how unreliable data controllers can monitor and supervise data leakage to third parties even if they are legally obligated to perform relevant duties.[145]

Therefore, we turn to the data subject itself and propose the right to authorize consent on data inferences. A listener-centric framework for personal data management geared to both economic effectiveness and human rights applies here. It offers the best solution for the conflict between free speech and privacy; that is, it empowers the users with the right to know, authorize, and revoke authorization to make data inferences. As users' access to and knowledge about data inferences is limited and the collective inference involving individual users' profiling is not specific enough for users to identify themselves, users must gain control over their data transmission through the right to authorize data inferences. As noted, data speech is robustly under the protection of the First Amendment unless and until the state affirmatively adopts audience-information collection rules.[146] It is important and necessary for data subjects to regain the power to authorize data inferences. Data inference authorization serves as a buffer for the acquisition of users' information.

Admittedly, harsh measurements might hinder the benefi-

[145] Emma Bowman, *After Data Breach Exposes 530 Million, Facebook Says It Will Not Notify Users*, NAT'L PUB. RADIO (April 9, 2021, 11:58 PM), https://www.npr.org/2021/04/09/986005820/after-data-breach-exposes-530-million-facebook-says-it-will-not-notify-users.

[146] Rauch, *supra* note 1, at 439 ("unless and until the state affirmatively adopts (appropriately drafted) audience-information collection rules, it has almost no power to limit political speakers' use of otherwise-lawfully obtained audience information in political Customized Speech.").

cial usage of data, as argued by economy-first scholars. For example, Tene and Polonetsky lament that a privacy-first approach would pose financial and technical burdens for both service providers, causing unmanageable consequences which would compromise beneficial usage of data.[147] Privacy policies of mainstream apps also exhibited a similar tone to this idea, as they gild their collection of users' behavioral information as a necessary step to improve their services and maximize users' experiences for internal operational purposes.[148]

We urge policymakers and service providers to recognize that the beneficial usage of data should not be manufactured upon the ignorance of users. Data controllers should not assume that the public would not donate their data or even digital privacies for the public good or service improvement.[149] Obtaining acknowledgement for anonymized data or data inferences from users is a prerequisite to eliminating the digital power asymmetry between platforms and the users, or the speakers and the listeners.

V. CONCLUSION

A decade ago, Paul Ohm observed that almost every single privacy statute and data protection regulation in the U.S. assumes that anonymization protects privacy, but this assumption becomes

[147] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 258 (2013).
[148] *See* INSTAGRAM, *Privacy Policy*, (July 26, 2022), https://privacycenter.instagram.com/policy ("We use information we have, including any information with special protections you choose to share, to provide and improve our Products. This includes personalizing features, content and recommendations, such as your Facebook Feed, Instagram feed, Stories and ads."); TIKTOK, *Privacy Policy for Younger Users*, (Jan. 2020), https://www.tiktok.com/legal/privacy-policy-for-younger-users?lang=en ("We share the information we collect with our corporate group and with service providers as necessary for them to perform a business purpose, professional service, or technology support function for us.").
[149] Yafit Lev-Aretz, Data Philanthropy, 70 HASTINGS L.J. 1491, 1500-1503 (2019). (exemplifying how private sector data is being donated for socially beneficial reuses).

malfunctional when easy re-identification makes PII-focused laws unproductive.[150] He warned us that the failure of anonymization disrupts privacy laws. Information decryption techniques never stopped developing ever since his warning. Ten years after, what we are facing is no longer a privacy issue. The capacity of data inferences has been broadened to exert a substantial and sometimes irreversible influence on both individual and societal levels.

This article refutes the idea that all data inferences implicate free speech rights by pointing out that devil data inferences with tangible harmful consequences do not deserve free speech protections. An excessive emphasis on free speech protections for data inferences can chill expression among technology users. This is the time when free speech protections backfire. To mitigate the potential risks of this backfire, we try to set forth a listener-centric approach by emphasizing listeners' free speech rights which data speech discussions have paid scant attention.

[150] Paul Ohm, *supra* note 63, at 1740.