Indiana Law Journal

Volume 100 | Issue 4 Article 5

6-2025

Unlocking Platform Data for Research

Niva Elkin-Koren
Tel-Aviv University, elkiniva@tauex.tau.ac.il

Maayan Perel
Netanya Academic College School of Law, maayanfilmar@gmail.com

Ohad Somech
Netanya Academic College School of Law, ohads2@gmail.com

Unlocking Platform Data for Research

NIVA ELKIN-KOREN,[*] MAAYAN PEREL[**] & OHAD SOMECH[***]

Digital platforms, which control unique access points to the rich data stored on their servers, have become a "living lab" of real-time information. Scientists and researchers increasingly use platform data for various purposes, such as training machine learning (ML) systems and Natural Language Processing (NLP) models, and for studying diverse fields such as medicine, humanities, and social sciences, including the influence of digital platforms on society. However, researchers increasingly encounter significant barriers when attempting to access platform data. Although platforms typically lack proprietary rights over the data itself, they exert strong control over its use by imposing digital locks and boilerplate contractual limitations. Faced with the legal risk of potential breach-of-contract lawsuits filed by well-funded platforms, researchers may simply opt to steer clear of platform data research.

This Article proposes private-law-centered solutions to overcome platform data lockout. First, researchers who access and use platform data without explicit permission should be able to contest breach-of-contract claims made against them by claiming copyright preemption. Platform data falls under copyright law, either because it is protected by copyright (such as user-generated-content) or because it constitutes basic "building blocks," such as users' digital data trails, which are specifically excluded from copyright protection. When platforms robustly ban any reproduction of data, they effectively benefit from quasi-copyright protection, through private ordering, albeit compromising fundamental copyright principles, including fair use. Their contractual claims should, therefore, be preempted by copyright law. Second, courts should facilitate platform data research by narrowly interpreting boilerplate contractual bans on data access. Third, nuisance law may further support platform data research by empowering researchers to demand the removal of technological barriers that hinder access to public, non-proprietary data.

Private law solutions to platform data lockout, however, do not grant researchers an affirmative right to use platform data for research. Legislative action of the type recently pursued by the European Union is required to establish such a right to research. This Article therefore concludes by examining regulatory approaches to platform data lockout, concluding that combining private law solutions with regulatory intervention offers the most effective means of adequately facilitating platform data research.

---

[*] Stewart and Judy Colton Professor in Legal Theory and Innovation, and the academic Director of the Shamgar Center for Digital Law and Innovation at Tel-Aviv University Faculty of Law.

[**] Assistant Professor, Netanya Academic College School of Law.

[***] Assistant Professor, Netanya Academic College School of Law.

This Research was supported by the Israel Science Foundation (grant No. 1870/21).

INTRODUCTION

In 2021, Facebook (now Meta) suspended the personal accounts of Laura Edelson and several other members of the New York University's (NYU) Cybersecurity for Democracy research team, denying them access to the platform.[1] At the time, Laura

1. *See* James Vincent, *Facebook Bans Academics Who Researched Ad Transparency and Misinformation on Facebook*, THE VERGE (Aug. 4, 2021, 7:08 AM), https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-

Edelson and her team were five months into an empirical study of the origin and dissemination of political ads and disinformation on Facebook.[2] To gather necessary data, they provided Facebook users with the option to download the "Ad Observer" plug-in, which automatically collects data on which political ads users encounter without collecting personal information that could identify the user.[3] In suspending the researchers' access, however, Facebook cited concerns about privacy protection of its users and its Terms of Service (ToS) prohibiting automatic data collection (scraping).[4]

Facebook proposed that the NYU team could continue their research through the platform's FORT research program, which offers limited information on political ad targeting that is both controlled and filtered by the platform itself.[5] Unsurprisingly, the researchers declined this offer, explaining that losing full access to Facebook's data dealt a death blow to the study.[6] Edelson accused Facebook of attempting to stifle her work, claiming the company aims to prevent "independent scrutiny of its platform." "Worst of all," Edelson added, "Facebook is using user privacy, a core belief that we have always put first in our work, as a pretext for doing this." She claimed that this incident highlighted the danger of granting Facebook "veto power over who is allowed to study them."[7]

The Ad-Observer study is but one example of a growing trend. Digital platforms increasingly impose restrictions on data access, even when the data is to be used solely for noncompeting scientific purposes.[8] This trend has not spared platforms that once cooperated with researchers. Twitter (now X), for example, which was once praised for its Application Programming Interface (API) that enabled researchers to access and search the platform's data,[9] decided in 2023 to shut its API down, adopting a fairly restrictive approach to data access.[10]

---

transparency-misinformation-nyu-ad-observatory-plug-in [https://perma.cc/M9RS-XUQD].

2. *See* Mariella Moon, *Facebook Disables Accounts of NYU Team Looking into Political Ad Targeting*, ENGADGET (Aug. 4, 2021), https://www.engadget.com/facebook-disables-accounts-nyu-ad-observatory-project-091040346.html [https://perma.cc/44HK-TTB7].

3. *Ad Observer*, NYU CYBERSECURITY FOR DEMOCRACY, https://adobserver.org [https://perma.cc/JR3W-UCR8].

4. *See* Ramishah Maruf, *Researchers Studying Facebook Misinformation Say They Were Deplatformed*, CNN, https://edition.cnn.com/2021/09/05/media/reliable-sources-facebook-researchers-deplatform/index.html [https://perma.cc/H5VF-8ACH] (Sept. 5, 2021, 5:36 PM) ("'We took these actions to stop unauthorized scraping and protect people's privacy' . . . Mike Clark, Product Management Director, said in a statement.").

5. *See* Vincent, *supra* note 1.

6. Maruf, *supra* note 4.

7. Vincent, *supra* note 1.

8. *See, e.g.*, Axel Bruns, *After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research*, 22 INFO., COMMC'N & SOC'Y 1544 (2019).

9. *See id.* at 1545–46; Justine Calma, *Twitter Just Closed the Book on Academic Research*, THE VERGE (May 31, 2023, 9:19 AM), https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research [https://perma.cc/J3YE-EUU4].

10. *See* Jenae Barnes, *Twitter Ends Its Free API: Here's Who Will Be Affected*, FORBES (Feb. 3, 2023, 6:08 PM), https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected [https://perma.cc/H54F-RD2A].

Platform data, as the Ad-Observer incident shows, is crucial both for studying platforms' behavior and for holding them accountable for that behavior. Indeed, the use of platform data has facilitated studies on platforms' role in social and political polarization,[11] discrimination in services and hiring decisions,[12] and biases in platforms' ranking and recommendation algorithms.[13] Beyond platform oversight and accountability, platform data is also invaluable in pursuing other scientific objectives, including training Natural Language Processing (NLP) and generative Artificial Intelligence (AI) models to develop new research methodologies,[14] and in studying body image, mental health, substance (ab)use, disaster risk reduction, and crisis management.[15]

From a legal-normative perspective, access to platform data can be conceptualized as part of the "right to research." This right, encompassing the ability to conduct research and obtain access to scientific findings, is essential for informed citizenship in a democratic society[16] and for fostering economic growth through entrepreneurship and innovation.[17] It is grounded in individuals' fundamental rights, including the right to "freedom of information and the public's right to information."[18] At the same time, it is justified by the public interest in advancing science for the benefit of humanity as a whole. The latter is reflected in the exclusion of data from copyright protection as well as in the copyright law's limitations and

---

11. *See, e.g.*, Abraham Israeli & Oren Tsur, *Free Speech or Free Hate Speech? Analyzing the Proliferation of Hate Speech in Parler*, *in* PROCEEDINGS OF THE FOURTH WORKSHOP ON ONLINE ABUSE AND HARMS (WOAH) 109 (2022).

12. *See, e.g.*, Sandvig v. Barr, 451 F. Supp. 3d 73 (D.D.C. 2020); Morgane Laouénan & Roland Rathelot, *Can Information Reduce Ethnic Discrimination? Evidence from Airbnb*, 14 AM. ECON. J. APPLIED ECON. 107 (2022); Benjamin Edelman, Michael Luca & Dan Svirsky, *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment*, 9 AM. ECON. J. APPLIED ECON. 1 (2017).

13. Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike & Nikhil Garg, *Fair Ranking: A Critical Review, Challenges, and Future Directions*, *in* FACCT '22: PROCEEDINGS OF THE 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1929, 1936 (2022).

14. *See, e.g.*, Jessamy Perriam, Andreas Birkbak & Andy Freeman, *Digital Methods in a Post-API Environment*, 23 INT'L J. SOC. RSCH. METHODOLOGY 277, 286 (2020).

15. *See, e.g.*, Camille Nebeker, Sarah E. Dunseath & Rubi Linares-Orozco, *A Retrospective Analysis of NIH-Funded Digital Health Research Using Social Media Platforms*, 6 DIGIT. HEALTH 1 (2020); David E. Alexander, *Social Media in Disaster Risk Reduction and Crisis Management*, 20 SCI. & ENG'G ETHICS 717 (2014).

16. Arjun Appadurai, *The Right to Research*, 4 GLOBALISATION, SOC'YS & EDUC. 167, 168 (2006).

17. Philip Barrett, Niels-Jakob Hansen, Jean-Marc Natal & Diaa Noureldin, *Why Basic Science Matters for Economic Growth*, IMF BLOG (Oct. 6, 2021), https://www.imf.org/en/Blogs/Articles/2021/10/06/blog-ch3-weo-why-basic-science-matters-for-economic-growth [https://perma.cc/UT2W-4TJC].

18. Sean Flynn, Christophe Geiger, João Pedro Quintais, Thomas Margoni, Matthew Sag, Lucie Guibault & Michael Carroll, *Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action*, 42 EUR. INTELL. PROP. REV. 393, 395 (2020); *see also, e.g.*, Appadurai, *supra* note 16, at 171, 175–77.

exceptions, specifically, the fair use doctrine, which permits reuses of copyrighted material for socially beneficial purposes such as research.[19]

Because platforms control unique access points to the rich data stored on their servers, and because that data is critical for conducting scientific studies in the digital era, access to platform data forms an integral part of the right to research. Accordingly, the prevalence of contractual and technological barriers to accessing platform data raises concerns regarding the ability of researchers to conduct independent, scientific, public interest research on digital platforms. These concerns have prompted various proposals for legal reform, each tackling a particular dimension of data access barriers.[20]

With few exceptions, platforms typically lack proprietary rights in the data on their servers. Intellectual Property (IP) laws exclude data from their protection specifically to fulfill their primary objective of fostering creation and innovation.[21] Nonetheless, platforms employ various contractual and technological measures to effectively restrict access to data. For instance, the majority of platforms' ToS include broad restrictions on manual and automatic data collection.[22] Additionally, platforms employ other, less explicit methods to contractually limit researchers' access to data, including terms prohibiting users from circumventing the platforms' technological barriers and mandating the use of real names when accessing and using the platform.[23]

Faced with the legal risk of a breach-of-contract lawsuit filed by a well-funded platform, researchers may opt to steer clear of platform data research. The contractual and technological barriers erected by platforms thus enable them to exert exclusive control over data to which they have no proprietary rights, thereby undermining the public interest in advancing scientific knowledge, in preventing the emergence of "information monopolies," in promoting democracy, and in holding platforms accountable.[24]

This Article contends that, properly applied, private law allows researchers to successfully defend against breach-of-contract claims that platforms bring against them. First, from an IP perspective, researchers should rely on the copyright preemption doctrine to contest such claims, which essentially reflect an illegitimate attempt to contract around copyright while disrupting copyright law's internal balance between exclusivity and access.[25] All types of platform data fall under copyright law, either because they are protected by copyright (such as user-generated content), or because they constitute basic "building blocks" intentionally excluded

---

19. *See* Flynn et al., *supra* note 18, at 393–96.

20. *See, e.g.*, Jacquellena Carrero, Note, *Access Granted: A First Amendment Theory of Reform of The CFAA Access Provision*, 120 COLUM. L. REV. 131 (2020); Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743 (2021); Benjamin L. W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147 (2021).

21. *See infra* Section I.C.

22. *See* Flynn et al., *supra* note 18, at 398.

23. *See, e.g.*, Sandvig v. Barr, 451 F. Supp. 3d 73, 76–77 (D.D.C. 2020) (discussing attempts to study race and gender discrimination in employment websites by using multiple fake accounts).

24. *See infra* Section III.B.

25. *See infra* Section III.A.

from copyright protection (such as users' digital data trails). Furthermore, because platforms actively strive to maintain exclusive control over the data they host by banning its reproduction, they effectively benefit from quasi-copyright protection through private ordering. Such attempts to appropriate platform data through boilerplate contract provisions undermine fundamental copyright principles, which promote the extraction of unprotected facts for research purposes.[26]

Second, contract law's rules of interpretation, as well as the unconscionability doctrine, further suggest that researchers should prevail in a breach-of-contract dispute with a platform. As a power-conferring instrument, contracts enable private actors to establish the terms of their interactions. However, these powers are not absolute; parties' ability to determine the rules of contract formation and interpretation is limited. Employing contracts to deny access to platform data for noncommercial, scientific purposes may exceed these powers, and thus, such contracts might be annulled. Narrowly interpreting contractual terms that robustly restrict access to platform data as inapplicable to access made for noncommercial, scientific purposes reflects researchers' reasonable expectations and advances the public interest. Contract law calls on courts to consider these outcomes when interpreting agreements, especially standard-form contracts like platforms' ToS.[27]

Third, nuisance law may additionally support platform data research by empowering researchers to demand the removal of technological barriers hindering access to public, non-proprietary data. When adjudicating instances of private landowners erecting physical barriers to access public resources, such as beaches and public lands, courts have found the barriers to constitute a nuisance and thus ordered their removal.[28] More recently, and in the context of intangible resources, the Ninth Circuit Court of Appeals has ordered LinkedIn to remove its technological barriers and allow a competing firm to access the public data on the platform.[29] The justification for similar injunctions is even greater when considering access to data for noncompeting, scientific purposes.[30]

Private law thus offers researchers an array of doctrines and principles to defend against potential breach-of-contract claims and even request the removal of digital locks on data access. Nevertheless, private law does not confer upon researchers an affirmative right to request and receive specific platform data for research. Such an affirmative right to research requires legislative action of the type recently taken by the European Union in its Digital Services Act.[31] Accordingly, we end this Article by discussing the advantages and shortcomings of a regulatory solution to platform data lockout and conclude that combining a bottom-up private law approach with top-down regulatory interventions might be the best approach to adequately securing a right to platform data research.

---

26. *Id.*

27. *See infra* Section III.B.

28. *See, e.g.*, Camfield v. United States, 167 U.S. 518 (1897).

29. *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1185, 1188 (9th Cir. 2022).

30. *See infra* Section III.B.3.

31. Council Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1 [hereinafter DSA].

The Article proceeds as follows: Part I provides the underlying foundations for the problem of platform data lockout. First, it describes the growing use of platform data in various research fields and demonstrates how critical access to platform data has become for scientists. Then, this Part maps the different components of platform data and explains why they hardly afford platforms with any proprietary rights. Part II turns to describe the problem of platform data lockout, critically analyzing the different reasons platforms assert for locking data on their servers. Part II then discusses the contractual and technological barriers that platforms deploy to dissuade third parties, including researchers, from accessing and utilizing platform data without permission, out of fear of facing lawsuits for breach-of-contract. Part III then delves into the major normative contribution of the Article, proposing three private law solutions to address the platform data lockout: the first relying on copyright law, the second on contracts law, and the third on the doctrine of nuisance. All three aim to enable researchers to effectively defend against breach-of-contract claims brought against them by platforms and to even demand the removal of technological barriers erected by platforms on data access. Finally, Part III addresses the limitations of these solutions, which fall short of creating an affirmative right to obtain specific platform data for research. The Article concludes by suggesting that the combination of private law solutions with regulatory interventions is the most effective approach to supporting a right to research platform data.

I. PLATFORM DATA AND ACADEMIC RESEARCH

A. Platform Data in Scientific Research

Platform data is a treasure chest for researchers—a "living lab" of unmeasurable and invaluable real-time data—providing scientists with opportunities to reach new frontiers of knowledge and understanding. Computer scientists use platform data to develop and train NLP and generative AI models.[32] In medical and health professions, platform data is used to study mental health, substance (ab)use, diagnosis, and weight and physical activity.[33] For researchers in the fields of humanities and social sciences, platform data is at the center of a "computational revolution," with an increasing number of scholars using platform data to research diverse phenomena "from the psychological underpinnings of human morality, to the influence of misinformation, to the factors that make some artists more successful than others."[34]

Platform data is also at the forefront of research on disaster risk reduction and crisis management.[35] As highlighted in a 2021 report by the Center for Strategic and International Studies (CSIS), platform data is a critical component in many disaster risk reduction strategies.[36] Researchers have already shown how real-time platform

32. *See, e.g.*, Perriam et al., *supra* note 14, at 286.

33. Nebeker et al., *supra* note 15, at 3.

34. Heidi Ledford, *Computing Humanity: How Facebook, Twitter and Other Sources Are Revolutionizing Social Science*, 582 NATURE 328, 328–29 (2020).

35. Alexander, *supra* note 15, at 720–24.

36. *See* Daniel F. Runde, Linnea Sandin & Arianna Kohan, *Disaster Risk Reduction Through Digital Transformation in the Western Hemisphere*, CTR. FOR STRATEGIC & INT'L

data supports models that provide early warning signs of food shortages[37] and materializes systemic risks in the banking system.[38] In 2013, researchers used data from (then) Twitter to develop the algorithm applied by responders to evacuate 10,000 flood victims during the Colorado flooding disaster.[39]

Finally, platform data is also fundamental in research on digital platforms' accountability. Examples include studies on platforms' political-ad-targeting practices;[40] social and political polarization;[41] online discrimination in employment and services;[42] "fake news" and disinformation;[43] and fairness and transparency in digital platforms' ranking and recommendation systems.[44]

Data-driven scientific research predates the emergence of digital platforms. Two related reasons account for platform data's unique significance. First, often the platform data relevant for scientific research is observed and collected solely by digital platforms,[45] giving them exclusive control of any access points to such data. Second, platform data's distinctive characteristics and its subsequent scientific significance can be encapsulated succinctly by the phrase "garbage in garbage out."[46] In data-driven research, the quality of research outcome is often as good as the

---

STUD. BRIEFS, Sept. 2021, at 6–8, https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/210927_Runde_Disaster_Risk_Reduction_1.pdf?VersionId=DuO8PB9MG2bojuT7MJKsdBbNYoMX7ZsF [https://perma.cc/MN3L-LHZC]; *see also* ORG. FOR ECON. CO-OPERATION & DEV., DATA-DRIVEN INNOVATION: BIG DATA FOR GROWTH AND WELL-BEING 31, 50, 395 (2015), https://www.oecd.org/en/publications/data-driven-innovation_9789264229358-en.html [https://perma.cc/3V2S-SFKW] ("Real-time analysis of a wide range of data generated through social media, mobile devices and physical sensors . . . provides a new opportunity for addressing complex social challenges, including in particular crisis prevention and disaster management.").

37. *See* ORG. FOR ECON. CO-OPERATION & DEV., *supra* note 36, at 28, 31 n.23.

38. Paola Cerchiello, Paolo Giudici & Giancarlo Nicola, *Twitter Data Models for Bank Risk Contagion*, 264 NEUROCOMPUTING 50, 50 (2017).

39. Megan Saltzman, *Social Media Mining: Can We Prevent the Apocalypse?*, 8 J. BIOSECURITY, BIODIVERSITY & BIODEFENSE L. 19, 19 (2017).

40. *See, e.g.*, Jeff Horwitz, *Facebook Seeks Shutdown of NYU Research Project into Political Ad Targeting*, WALL ST. J., https://www.wsj.com/tech/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533 [https://perma.cc/EKT4-SV3S] (Oct. 23, 2020, 8:59 PM).

41. *See, e.g.*, Israeli & Tsur, *supra* note 11.

42. *See, e.g.*, Sandvig v. Barr, 451 F. Supp. 3d 73 (D.D.C. 2020); Laouénan & Rathelot, *supra* note 12; Edelman et al., *supra* note 12.

43. *See, e.g.*, Craig Timberg & Elizabeth Dwoskin, *Facebook Takes Down Data and Thousands of Posts, Obscuring Reach of Russian Disinformation*, WASH. POST (Oct. 12, 2017) https://www.washingtonpost.com/news/the-switch/wp/2017/10/12/facebook-takes-down-data-and-thousands-of-posts-obscuring-reach-of-russian-disinformation/ [https://perma.cc/3UWN-WMNA].

44. *See, e.g.*, Patro et al., *supra* note 13.

45. *See infra* Section I.B.

46. JOHN DAINTITH & EDMUND WRIGHT, *Garbage in Garbage out (GIGO)*, OXFORD DICTIONARY OF COMPUTING (6th ed. 2008), https://www.oxfordreference.com/display/10.1093/acref/9780199234004.001.0001/acref-9780199234004-e-6364 [https://perma.cc/CLZ3-KEMJ] ("[A]ll input, however absurd, will be processed according to a program's algorithms . . . .").

quality of input data. Low-quality data can lead to inaccurate, biased, and misleading research outcomes. Assessed along the dimensions of the four Vs,[47] platform data contains an (almost) infinite amount of real-time and historical data (and metadata) pertaining to the thoughts, actions, opinions, and concerns of users, advertisers, and political and social actors from around the world. That is, information pertaining to people of diverse ethnic and socio-economic status, including "vulnerable populations that are traditionally difficult to reach,"[48] engaged in political, social, commercial, and personal interactions. Few, if any, databases can therefore match the volume, velocity, or variety of platform data.

B. Unpacking "Platform Data"

The term "data" is part of everyday language, often used to describe "information,"[49] which is "facts about a situation, person, [or] event"[50] recoded in a human or machine-readable way.[51] Data is separate from the particular way in which it is represented.[52] Thus, while the scoring of a touchdown may be represented in multiple ways (e.g., a video, an audio recording, or a particular sequence of words), we use the term "data" to refer to the underlying fact that a touchdown was scored—a fact independent from any particular way in which the information was observed, recorded, or communicated.

Data's properties differ from those of tangible resources. First, data is non-rivalrous, with multiple agents able to simultaneously use the same data (or dataset) for various purposes without diminishing its use-value to others.[53] Data can also be collected by multiple sources. A car's geolocation, for example, can be determined by the car's computer, the driver's smartphone, and the cameras installed on traffic lights.[54]

These properties of data might suggest that access to any particular dataset is of little importance. However, as Daniel Rubinfeld and Michal Gal pointed out, "unique access points to unique data may lead to situations in which the data cannot be easily replicated."[55] Delayed or partial access, as well as an inability to integrate multiple sources or verify the data's accuracy, would all reduce the data's quality, place its

---

47. Data's quality is often measured by the so-called four Vs: volume—the amount of data; velocity—the speed in which new data is integrated into the database; variety—the number of sources and time periods from which the data was collected; and value—the data's accuracy. *See, e.g.*, Mark Lycett, '*Datafication': Making Sense of (Big) Data in a Complex World*, 22 EUR. J. INFO. SYS. 381, 381–82 (2013).

48. Nebeker et al., *supra* note 15, at 1.

49. Lothar Determann, *No One Owns Data*, 70 HASTINGS L.J. 1, 6 (2018).

50. *See, e.g.*, *Information*, CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/dictionary/english/information [https://perma.cc/94R6-U7UH].

51. *See* Josef Drexl, *Designing Competitive Markets for Industrial Data*, 8 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 261, 263, 273 (2017).

52. *See id.*

53. That is, opposed to data's economic value. *See infra* Section II.A.

54. *See* Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 9 ARIZ. L. REV. 339, 350–51 (2017).

55. *Id.* at 351.

recipients at a disadvantage, and may even corrupt the outcome of any subsequent analysis.

Platform data includes raw and processed data,[56] personal and nonpersonal information,[57] and information pertaining to multiple subject matters, such as one's medical conditions, financial circumstances, political allegiance, and religious beliefs.[58]

Platform data may be generated by users, advertisers, the platform itself, or any combination of the three. Some types of platform data are published to and observable by the public at large, while the publication and observability of other types of data are restricted to registered (logged-in) users who accepted the platform's contractual terms: paying advertisers or the platform's employees alone. Finally, most platforms restrict others' ability to independently record platform data (e.g., via bots or third-party extensions).[59]

For the purpose of our discussion, we delineate among three types of platform data. We do not intend for these distinctions to be rigid, and boundary cases and overlaps exist between categories. Instead, the purpose of this categorization is to facilitate the discussion on access to platform data within the current framework of data legal regimes, as well as the economic, social, and other fairness considerations pertaining to data access. The first type of data, User Produced Data (UPD), refers to information generated, observed, and recorded by users who subsequently publish it on the platform, such as the content of a user's post. UPD is often made available to the general public, although some platforms offer users the option to restrict who may access this information (e.g., platform members or other users within the user's network).

The second type of data is Co-Produced Data (CPD), which is generated by users or advertisers, but observed, measured, and recorded by the platform. CPD typically involves the platforms' observation, measurement, and recording of users' and advertisers' behavior, such as the frequency of a user logging into the platform each day, the on-screen location of the user's pointer, or the speed a user presses the "like" button.[60] CPD may exist in aggregated or individualized forms, and it can be

---

56. See *id.* at 350–52, for an example of this distinction.

57. See Council Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on The Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016, art. 4(1) O.J. (L 119) 33 [hereinafter GDPR], for this classification of data ("'[P]ersonal data' means any information relating to an identified or identifiable natural person."). See RESTATEMENT (SECOND) OF TORTS § 652D (AM. L. INST. 1977), for a narrower approach ("One who gives publicity to a matter concerning the private life of another is subject to liability to the other for invasion of his privacy . . . .").

58. *See, e.g.*, Genetic Information Nondiscrimination Act (GINA) of 2008, 42 U.S.C. §§ 2000ff, ff-1 to -11; Financial Services Modernization Act of 1999, 15 U.S.C. §§ 6801–09, 6821–27; Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. §§ 1320d, d-1 to -9; Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g.

59. *See* Casey Fiesler, Nathan Beard & Brian C. Keegan, *No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service*, 14 PROC. OF THE INT'L AAAI CONF. ON WEB & SOC. MEDIA 187, 190–91 (2020).

60. Specifically, Facebook's (Meta's) "like" button not only measures who and how many users "liked" a certain Facebook post, but it also creates a measuring unit for posts'

processed or raw. However, access to CPD is often restricted to platforms' employees and paying customers.

Finally, some platform data is Platform Produced Data (PPD). PPD is information that is generated, observed, and recorded by the platform. PPD often does implicate human behavior but is nevertheless removed from it. For example, the number of people in a user's network (or the average network) is the product of human behavior (the sending and accepting of invitations to join the network). Yet rather than an observation (or recording) of human behavior, it is the accumulated analysis of such behavior over time. The relation between PPD and human behavior, therefore, is more similar to that between the latter and the number of people living in the United States or the Earth's surface temperature. PPD may be aggregated or individualized, processed or raw, with access usually restricted to platforms' employees and paying customers.[61]

C. Rights in Platform Data: Public Domain and Private Ownership

Many types of platform data, including those that are observable by the public at large, are kept locked behind platforms' boilerplate ToS and technological barriers.[62] Platforms resort to such private ordering measures of data control because, as discussed below, they have little, if any, proprietary rights in the data itself.

IP laws explicitly exclude data from their protectable subject matter.[63] The exclusion of data from IP protection follows the well-known idea/expression distinction, according to which IP protection applies to particular expressions of ideas, but does not extend to the idea itself or to any "procedure, process, system, method of operation, concept, principle, or discovery."[64] From a normative (policy) perspective, IP laws are commonly justified on the utilitarian grounds of incentivizing the creation of original and innovative works by granting exclusive rights to their creators.[65] Achieving this objective requires IP law to strike a delicate balance: Structuring IP rights too narrowly provides insufficient incentives; structuring them too broadly impedes the creation of subsequent works.[66]

---

popularity. And though one could post the words "like" when commenting on a post he "liked," this is unlikely to have the same public (objective) meaning as the pressing of the "like" button. Digital platforms, to be sure, not only facilitate the measurement of information, but they also restrict it. For example, by limiting the interoperability of third-party extensions and applications, Facebook prevents the emergence of alternative ways to observe and measure users' reactions. *See* Thomas E. Kadri, *Digital Gatekeepers*, 99 TEX. L. REV. 951, 974–75 (2021).

61. *See, e.g.*, Rory Van Loo, *Privacy Pretext*, 108 CORNELL L. REV. 1, 22 (2022) (discussing Facebook's systematic monitoring and collection of data on "which apps were both (1) growing in popularity and (2) offering competing services").

62. *See infra* Part II.

63. Determann, *supra* note 49, at 11–12.

64. 17 U.S.C. § 102(b).

65. *See, e.g.*, Niva Elkin-Koren, *Copyright Policy and the Limits of Freedom of Contract*, 12 BERKLEY TECH. L.J. 93, 98–99 (1997).

66. *Id.*; Neil Weinstock Netanel, *Copyright and a Democratic Civil Society*, 106 YALE L.J. 283, 285 (1996) ("If copyright extends too broadly, copyright owners will be able to exert censorial control over critical uses of existing works or may extract monopoly rents for access,

IP laws use various tools to address this challenge. Most relevant to our discussion is the exclusion of certain subject matters from the protection of IP laws. Data is one such subject matter. As a building block for much innovative and creative works, granting rights in data would be counterproductive vis-à-vis IP law's objective of encouraging the production of innovative and creative works, and there is "no known 'data property statute' in any country."[67]

IP laws further balance the need to incentivize current and future works by allowing certain uses of protected works. In particular, to limit the extent to which IP rights impede subsequent works, the Copyright Act allows for the "use of a copyrighted work . . . for purposes such as criticism, . . . teaching[,] . . . scholarship, or research," especially when these are made for noncommercial or "nonprofit educational purposes."[68]

Some legal rules may nevertheless provide indirect rights in data. Most relevant to our discussion are rights in compilations of data (databases). Though copyright law does provide some rights in databases, these only apply to databases that demonstrate sufficient creativity and originality in the compilation of the data.[69] Thus, a database of UPD using generic categories is unlikely to be copyrightable, while a database of CPD, in which originality was put into the selection and categorization of the data, might be. Even then, rights in databases do not extend to the underlying pieces of information. Instead, they apply only against the copying of the entire database (or substantial parts of it) and remain subject to the fair use doctrine.[70]

Misappropriation laws provide another layer of protection for databases. Unlike copyright laws, the purpose of misappropriation laws is to protect investments in the collection of information against freeriding, and their application is therefore not subject to the creativity or originality requirements of copyright laws.[71] Nevertheless, misappropriation laws are limited in scope, applying to investments made deliberately for the creation and configuration of the databases, as opposed to databases that are mere by-products of other investments.[72] Moreover, and similar to the copyright protection of databases, misappropriation laws typically only protect against the "wholesale copying of the database or substantial parts of it, typically where freeriding could have a noticeable impact on investments and competition."[73]

---

thereby chilling discourse and cultural development.").

67. Determann, *supra* note 49, at 11.

68. 17 U.S.C. § 107.

69. *See* Determann, *supra* note 49, at 19; Aziz Z. Huq, *The Public Trust in Data*, 110 GEO. L.J. 333, 387–88 (2021).

70. *See* Drexl, *supra* note 51, at 267–68. In the EU, the Database Directive similarly only protects databases that exhibit creativity in the selection or arrangement of the data or whose creation required substantial investments (as opposed to the creation of the underlying data). Council Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, 1996 O.J. (L 77/20); *see also* Drexl, *supra*, at 267–68.

71. *See* Determann, *supra* note 49, at 21.

72. *Id.* at 22.

73. *Id.*; *see* ML Genius Holdings LLC v. Google LLC, No. 20-3113, 2022 WL 710744, at *2 (2d Cir. Mar. 10, 2022), for an example where some courts have further limited the applicability of misappropriation law (finding Genius's misappropriation claim to be

Indirect protection of data and databases may also be grounded in trade-secret laws. Trade-secret protection applies to nonpublic information, whose secrecy is necessary for its economic value, and where reasonable efforts were made to keep it secret.[74] Thus, trade-secret protection is unlikely to apply to publicly available data, such as UPD,[75] but may apply to nonpublic platform data, including raw data pertaining to in-platform advertisement and users' (online) behavior.

Finally, although privacy laws typically pertain to the use of data rather than its ownership, they may support indirect (*de facto*) rights in data constituting personal information. In the EU, the General Data Protection Regulation (GDPR) limits the transfer of personal information outside the EU,[76] requiring platforms to restrict American and other non-EU entities from accessing platform data. From a U.S. perspective, the effect of such limitations is practically similar to that of a proprietary (*in rem*) right to exclude. In the United States, restrictions on the use and transfer of data are content-dependent and extend to information such as medical and financial information.[77] Though these limitations will often be inapplicable to platform data, to the extent that they are, they may have a similar effect to the granting of *de facto* property rights in favor of the data-holding entity.[78]

II. TECHNOLOGICAL AND LEGAL BARRIERS TO ACCESSING DATA

Platforms hold quite limited propriety rights in the data they collect, process, or host. Most types of data, especially UPD and users' data trails, do not confer any ownership interest on platforms. Yet, all these types of data are still either generated, measured, collected, hosted, or stored by and on platforms' servers. Facebook generates the order and selection of content depicted in users' "news feeds";[79]

---

preempted by copyright law); *see also In re* BRCAl- & BRCA2-Based Hereditary Cancer Test Pat. Litig., 774 F.3d 755 (Fed. Cir. 2014) (rejecting Myriad's claim of Ambry's freeriding on its deliberate investment in creating a database of genetic invariant information).

74. 18 U.S.C. § 1839(3)(A)–(B); CAL. CIV. CODE §§ 3426.1(d), 3426.11 (West 2024).

75. *But see* Compulife Software, Inc. v. Newman, 111 F.4th 1147, 1160–62 (11th Cir. 2024) (granting trade-secret protection to a database of insurance quotes despite such data, in principle, being publicly available).

76. *See* GDPR, *supra* note 57, at art. 44–50.

77. *See, e.g.*, Genetic Information Nondiscrimination Act (GINA) of 2008, 42 U.S.C. §§ 2000ff, ff-1 to -11; Financial Services Modernization Act of 1999, 15 U.S.C. §§ 6801–09, 6821–27; Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. §§ 1320d, d-1 to -9.

78. HIPAA, for example, restricts the transfer and use of medical information. In principle, HIPAA provides patients with the right to receive their medical records at a price reflecting the data-holder's cost. *See* Privacy of Individually Identifiable Health Information, 45 C.F.R. §§ 164.502–512, 164.524. But in *Ciox Health, LLC. v. Azar*, 435 F. Supp. 3d 30 (D.D.C. 2020), the court refused to apply the price cap to patients' request to transfer their information to a third party (e.g., their lawyer or insurance company). To the extent patients' records can only be obtained from a single data-holder, HIPAA provides the latter with the right to exclude others from the data coupled with an unrestricted right to charge fees for accessing it, a combination that, to an extent, resembles the rights provided to owners.

79. *How Feed Works*, FACEBOOK, https://www.facebook.com/help/1155510281178725 [https://perma.cc/9S87-5NFJ].

YouTube measures users' preference and processes them by its recommendation algorithm;[80] Google collects data about users' search queries;[81] X (formerly Twitter) hosts users' tweets, and all of these are stored on the platforms' servers.[82] As users' content, personal data, and activities predominantly take place on their servers, platforms have the ability to technically block access to and impose contractual restrictions on all such data. Through the implementation of various private ordering mechanisms—particularly contractual restrictions and technological barriers—platforms hinder third parties, including researchers, from obtaining access to platform data and attempt to cast doubt on its legality.

In the following discussion, we turn to explain why platforms are reluctant to share their data with researchers and then describe the various measures platforms deploy to restrict platform data research.

A. Platform Data Lockout—The Driving Forces

The commercial interest of platforms is the key force driving platform data lockout. All types of platform data can be (and are) used for commercial purposes, including product design, pricing, marketing, and targeted advertising.[83] Platform data is also used in machine learning (ML) and Natural Language Processing (NLP).[84] For example, the content of users' tweets is used to train NLP models,[85] and the location of users' pointer can be used as raw data to train ML algorithms detecting market trends and consumer preferences.[86]

Because data is non-rivalrous, multiple firms could, in principle, use a single database of platform data to train their algorithms without undermining the ability of other firms to do the same. In practice, however, this is rarely the case. Instead, digital platforms seek to restrict access to and use of data hosted on their facilities. One reason they do so is to gain a competitive advantage, using the data to advance

---

80. Cristos Goodrow, *On YouTube's Recommendation System*, YOUTUBE OFF. BLOG (Sept. 15, 2021), https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/ [https://perma.cc/ZQQ7-LRRD].

81. *Privacy Policy*, GOOGLE, https://policies.google.com/privacy?hl=en [https://perma.cc/W7F8-KP8B].

82. *See, e.g.*, TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 18 (2018) (defining platforms as websites that "host, organize, and circulate users' shared content or social interactions for them").

83. *See, e.g.*, Nikolas Guggenberger, *Essential Platforms*, 24 STAN. TECH. L. REV. 237, 260–61, 265 (2021).

84. *See* Lemley & Casey, *supra* note 20, at 744–45, 753–54.

85. Limarc Ambalina, *20+ Twitter Datasets for NLP and Machine Learning Projects*, LINKEDIN (Oct. 26, 2020), https://www.linkedin.com/pulse/20-twitter-datasets-nlp-machine-learning-projects-limarc-ambalina [https://perma.cc/6XSJ-VYBG].

86. Amazon, for example, collects user-generated data to "decide how to price an item, which features to copy or whether to enter a product segment based on its earning potential." Dana Mattioli, *Amazon Scooped Up Data from Its Own Sellers to Launch Competing Products*, WALL ST. J., https://www.wsj.com/articles/amazon-scooped-up-data-from-its-own-sellers-to-launch-competing-products-11587650015 [https://perma.cc/6YD8-MSD8] (Apr. 23, 2020, 9:51 PM).

their business interest and preventing competitors from doing the same. Another reason is to inflate the price of data by making it (artificially) scarce and selling it to third parties of their choosing. With data being the primary revenue generating asset for data-driven firms like Google,[87] platforms hold a significant economic interest in restricting competitors' access to data and increasing its market price.

Platforms also have a significant reputational interest in limiting access to data. Platforms may restrict access to enhance their reputation among users as champions of users' privacy,[88] though the merit of this strategy may be questionable,[89] especially when considering that platforms sell data to any third party willing to pay their asking price. Perhaps more importantly, then, platforms restrict access to data to deny others the ability to criticize and hold them accountable. Evidence of platforms' discriminatory (or discrimination-enabling) activities, for example, can often only be attained by accessing and analyzing platform data. Because revealing this information is harmful to the platform's reputation and may result in unwanted regulatory attention,[90] platforms have a strong reputational interest in controlling who may access platform data and for what purposes.[91]

Finally, and alongside their direct business interests, platforms limit data access to minimize potential legal risks and avoid liability for violating users' privacy, data protection laws (mainly the GDPR), or security regulations.[92] The aftermath of the Facebook/Cambridge Analytica scandal, which involved the misuse of millions of Facebook users' personal data acquired by researchers through Facebook's API, is a prominent example.[93] A five-billion-dollar fine inflicted on Facebook and new data privacy and security requirements set by the Federal Trade Commission (FTC)[94] led online platforms to restrict access to their APIs by any third parties in order to prevent abusive treatment of sensitive data. This was evident in Facebook's decision to shut down the accounts of team members of NYU's Cybersecurity for Democracy Project, disable the apps and Facebook pages associated with the project and its operators, and threaten legal action against them for accessing users' data in violation of Facebook's ToS.[95] As mentioned, the NYU research team was studying

87. *See* Huq, *supra* note 69, at 346 ("More than four-fifths of Alphabet's revenue derives from the sale of advertisements targeted using this data.").

88. *See* Deen Freelon, *Computational Research in the Post-API Age*, 35 POL. COMM. 665, 667 (2018).

89. *See generally* Van Loo, *supra* note 61.

90. *See, e.g.*, Bruns, *supra* note 8, at 1558.

91. *See* Thomas E. Kadri, *Platforms as Blackacres*, 68 UCLA L. REV. 1184, 1189 (2022).

92. *See* Bruns, *supra* note 8, at 1547–51, 1558 (discussing platform-imposed restrictions on access to data in the aftermath of the Cambridge analytics scandal).

93. Mark Townsend, *Facebook-Cambridge Analytica Data Breach Lawsuit Ends in 11th Hour Settlement*, THE GUARDIAN, (Aug. 27, 2022, 12:16 PM), www.theguardian.com/technology/2022/aug/27/facebook-cambridge-analytica-data-breach-lawsuit-ends-in-11th-hour-settlement [https://perma.cc/AR4P-XK8E].

94. FTC, STATEMENT OF CHAIRMAN JOE SIMONS AND COMMISSIONERS NOAH JOSHUA PHILLIPS AND CHRISTINE S. WILSON *IN RE FACEBOOK, INC.* (2019), https://www.ftc.gov/system/files/documents/public_statements/1536946/092_3184_facebook_majority_statement_7-24-19.pdf [https://perma.cc/SJ5H-YCAQ].

95. Meghan Bobrowsky, *Facebook Disables Access for NYU Research into Political-Ad Targeting*, WALL ST. J. (Aug. 4, 2021, 5:54 PM), https://www.wsj.com/articles/facebook-cuts-

misinformation in political ads on Facebook and, to collect data about the political ads displayed to Facebook users, used an Ad-Observer browser extension that allowed the sharing of data with the NYU research team. To justify the shutdown, Facebook claimed it had to comply with FTC rules because the researchers did not have permission from Facebook users to scrape the information.[96] While the FTC denied Facebook's assertions, and even sent a response letter to Facebook CEO Mark Zuckerberg, this example shows how platforms attempt to exploit regulatory obligations to safeguard their own, narrow business interests, even at the price of hampering socially beneficial research.

B. Contractual Limitations on Data Access

The key legal mechanism for locking platform data is contract law. Platforms regularly deploy various contractual limitations to restrict researchers' access to platform data via their ToS, privacy policies, or community guidelines.[97] These different contractual instruments are boilerplate contracts,[98] offered to users on a take it or leave it basis, and robustly binding all platform users, regardless of the purpose for which they use or access the platform. Deep power imbalances characterize the contractual relationship between researchers, users, and platforms.[99] This was observed by the court in *X Corp. v. Bright Data Ltd.*,[100] which rejected X's breach of contract claim against Bright Data for scraping publicly available data:

> We are not concerned here with an arm's length contract between two sophisticated parties in which one or the other adjusts their rights and privileges under federal copyright law. We are instead concerned with a massive regime of adhesive terms imposed by X Corp. that stands to fundamentally alter the rights and privileges of the world at large (or at least hundreds of millions of alleged X users).[101]

As we mentioned earlier, platforms control unique access points to data that may not be easily replicated.[102] Furthermore, users are not only denied a voice in shaping

---

off-access-for-nyu-research-into-political-ad-targeting-11628052204 [https://perma.cc/32DB-FM9Q].

96. Lois Anne DeLong, *Facebook Disables Ad Observatory; Academicians and Journalists Fire Back*, NYU CTR. FOR CYBERSECURITY (Aug. 21, 2021), https://cyber.nyu.edu/2021/08/21/facebook-disables-ad-observatory-academicians-and-journalists-fire-back/ [https://perma.cc/H9EJ-TXQE].

97. *See* Fiesler et al., *supra* note 59, at 187–89.

98. *See* Friedrich Kessler, *Contracts of Adhesion—Some Thoughts About Freedom of Contract*, 43 COLUM. L. REV. 629, 632 (1943). *See generally* Todd D. Rakoff, *Contracts of Adhesion: An Essay in Reconstruction*, 96 HARV. L. REV. 1173 (1983) (discussing the use of contracts of adhesion in business practices).

99. *See* Niva Elkin-Koren, Giovanni De Gregario & Maayan Perel, *Social Media as Contractual Networks: A Bottom Up Check on Content Moderation*, 107 IOWA L. REV. 987, 1023–24 (2022).

100. 733 F. Supp. 3d 832, 850 (N.D. Cal. 2024).

101. *Id.*

102. *See supra* note 55 and accompanying text.

these restrictions before accepting the contract terms, but they also have no assurances that these restrictions will remain in force without amendment from their initial form. Indeed, platforms often reserve absolute discretion to amend their ToS as they see fit, even without providing prior notice to users.[103] Hence, researchers may face serious legal challenges when attempting to access platform data, fearing potential legal liability while also having to declare to their research institutions and funders that data collection complies with legal and ethical requirements.

The threat of a legal suit from heavily funded digital platforms could create a significant barrier to access for researchers. In 2020, AlgorithmWatch initiated a project to monitor Instagram's newsfeed, which depended on data provided by volunteers who agreed to install a browser add-on that scraped their newsfeeds.[104] In early 2021, Facebook had contacted AlgorithmWatch, claiming that the project violated its ToS, which state that one "may not access or collect data from [Facebook's products] using automated means."[105] Facebook further claimed that AlgorithmWatch was violating the GDPR. As a result, AlgorithmWatch decided to terminate its project, explaining on its website that "an organization the size of AlgorithmWatch cannot risk going to court against a company valued at one trillion dollars."[106]

Many platforms, like Facebook, include straightforward prohibitions on data collection in their various contracts with users.[107] For instance, under LinkedIn's user agreement, users are not allowed to "[d]evelop, support or use software, devices, scripts, robots or any other means or processes . . . to scrape or copy the Services, including profiles and other data from the Services."[108] Data scraping refers to "the process of extracting and combining content of interest from the Web in a systematic way." Scraping "take[s] raw data in the form of HTML code from sites and convert[s] it into a usable structured format."[109] Robust bans on data scraping could be asserted against researchers who access and reproduce platform data, including public data, for research purposes. For instance, in a recent lawsuit filed by X against the Center for Countering Digital Hate (CCDH),[110] a nonprofit organization that conducted research on the dissemination of hateful content on social media, X alleged that CCDH had intentionally and unlawfully scraped data from X, thereby violating its ToS.[111] Rejecting X's claim, "[t]he Court notes, too, that X Corp.'s

---

103. Elkin-Koren et al., *supra* note 99, at 1037.

104. Nicolas Kayser-Bril, *AlgorithmWatch Forced to Shut Down Instagram Monitoring Project After Threats from Facebook*, ALGORITHMWATCH (Aug. 13, 2021), https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/ [https://perma.cc/TC2C-AQF5].

105. *Id.* (alteration in original).

106. *Id.*

107. *See* Fiesler et al., *supra* note 59, at 190–91.

108. *User Agreement: § 8.2(2)*, LINKEDIN, www.linkedin.com/legal/user-agreement [https://perma.cc/HGA5-WYYN] (Nov. 20, 2024).

109. Domenico Trezza, *To Scrape or Not to Scrape, This Is Dilemma. The Post-API Scenario and Implications on Digital Research*, FRONTIERS SOCIO., Mar. 2023, at 1, 2.

110. X Corp. v. Ctr. for Countering Digit. Hate, Inc., 724 F. Supp. 3d 948, 980–82 (N.D. Cal. 2024).

111. *Id.* at 968 ("X Corp. further alleges . . . that CCDH U.S. violated the ToS by scraping

motivation in bringing this case is evident. X Corp. has brought this case in order to punish CCDH for CCDH publications that criticized X Corp.—and perhaps in order to dissuade others who might wish to engage in such criticism."[112]

Data scraping is, of course, a vital research methodology as it allows researchers to independently extract their own database instead of relying on potentially incomplete or biased databases provided by platforms.[113] Moreover, data scraping is critical for dynamic exploratory research in which the explicit research questions and methods may not always be defined at the outset.[114] In such studies, research questions are derived from the data collected, and, therefore, they are especially dependent on accessing and collecting as much data as possible. Restricting researchers' ability to scrape data independently thus hinders scientific research.

Other direct restrictions on data access include contractual bans on conducting surveys,[115] or limitations on the dissemination of studies conducted using platform data. TikTok, as an example, has recently amended its research API ToS,[116] requiring academics to provide advance notice of their forthcoming research, subject their work to pre-publication review, and delete certain data once it has been used.[117]

Platforms' contractual terms may also impede access to data more implicitly. One example is the contractual requirement that users use their real names and open only one account.[118] Such a restriction is a major barrier for researchers seeking to study discrimination.[119] In *Sandvig*, for instance, researchers sought to study race and gender discrimination in employment websites by creating multiple fake accounts in violation of the websites' ToS, which prohibits misrepresentation. Although the district court refused to hold the researchers in violation of the Computer Fraud and

---

the X platform . . . .").

112. *Id.* at 981.

113. For instance, during the recent COVID-19 pandemic, Apple and Google launched a privacy preserving API for COVID-19 tracing apps, which did not support the collection of location data. *See* Mark Scott, Elisa Braun, Janosch Delcker & Vincent Manancourt, *How Google and Apple Outflanked Governments in the Race to Build Coronavirus Apps*, POLITICO (May 15, 2020, 5:25 AM) https://www.politico.eu/article/google-apple-coronavirus-app-privacy-uk-france-germany/ [https://perma.cc/RD9N-KARX].

114. *See* Bryan Weichelt, Priya Nambisan, Rick Burke & Casper Bendixsen, *Finding the Edges of Problems: Social Media as an Exploratory Research*, 25 J. AGROMEDICINE 423, 425 (2020); Richard Swedberg, *Exploratory Research*, *in* THE PRODUCTION OF KNOWLEDGE: ENHANCING PROGRESS IN SOCIAL SCIENCE 30–31 (Colin Elman, John Gerring & James Mahoney eds., 2020).

115. *See, e.g.*, *Terms of Service*, META, https://www.facebook.com/legal/terms/ [https://perma.cc/JN6N-2ZVR] (Jan. 1, 2025).

116. *TikTok Research Tools Terms of Service*, TIKTOK (Oct. 25, 2024), https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en [https://perma.cc/BK4U-BQRE].

117. *Id.*

118. *See, e.g.*, *Names Allowed on Facebook*, FACEBOOK HELP CTR. https://www.facebook.com/help/229715077154790 [https://perma.cc/R9NY-28KS].

119. *See, e.g.*, Sandvig v. Barr, 451 F. Supp. 3d 73, 76–77 (D.D.C. 2020).

Abuse Act (CFAA),[120] it did state that such misrepresentation might constitute a violation of the website's ToS.[121]

Other examples of implicit restrictions on data access include prohibitions on technical overburdening of the platform[122] and restrictions on data portability[123] and data transfer.[124]

Platforms may further use their ToS agreements to shield the technological barriers they embed in their systems for the purpose of limiting data access. On the technological side, platforms typically use username and user-created password authentication gates to prevent nonusers' access to platform data.[125] However, platforms also use more sophisticated technological tools to prevent others from accessing their data, leading to what the Ninth Circuit aptly called "technological gamesmanship."[126]

Craigslist and Facebook, for example, use internet protocol address (IP address) blockers to prevent third parties from scraping the data on their servers.[127] In *hiQ Labs, Inc. v. LinkedIn Corp.*,[128] the court provided a detailed overview of the

---

120. *See* 18 U.S.C. § 1030(a) ("Whoever . . . intentionally access a computer without authorization or exceeds authorized access and thereby obtains . . . information from any protected computer . . . shall be punished . . . .").

121. *See Sandvig*, 451 F. Supp. 3d, at 77, 80.

122. *See Terms of Service 3.2(2)*, META (Jan. 1, 2025), https://www.facebook.com/terms.php [https://perma.cc/JN6N-2ZVR]; *User Agreement 8.2(3)*, LINKEDIN (Nov. 20, 2024), https://www.linkedin.com/legal/user-agreement#dos [https://perma.cc/HGA5-WYYN]; Ryanair DAC v. Booking Holdings Inc., 636 F. Supp. 3d 490, 503 (D. Del. 2022) ("Ryanair alleges that the actions of the defendants and/or their agents 'greatly increase[ ] the quantities of queries on the Ryanair Website,' 'impair[ ] the . . . availability and/or usability' of the Ryanair website, and cause the website's response times to deteriorate." (alterations in original) (citations omitted)); X Corp. v. Ctr. for Countering Digit. Hate, Inc., 724 F. Supp. 3d 948, 979–80 (N.D. Cal. 2024) (same); X Corp. v. Bright Data Ltd., 733 F. Supp. 3d 832, 842 (N.D. Cal. 2024) ("Absent allegation, it cannot be assumed that Bright Data or its customers sending requests to X Corp.'s servers with a scraper is inherently burdensome, or inherently more burdensome than an X user sending requests to X Corp[] . . . .").

123. Steve Satterfield, *Transfer Your Facebook Posts and Notes with Our Expanded Data Portability Tool*, META (Apr. 19, 2021), https://about.fb.com/news/2021/04/transfer-your-facebook-posts-and-notes-with-our-expanded-data-portability-tool/ [https://perma.cc/25TL-RGP4].

124. *Id.*

125. *See* Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596, 1664 (2003).

126. Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016).

127. *E.g.*, Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962, 966–67, 969–70 (N.D. Cal. 2013); Meta Platforms, Inc. v. Bright Data Ltd., No. 23-cv-00077-EMC, 2024 WL 251406, at *1 (N.D. Cal. Jan. 23, 2024) ("Meta has also developed a variety of technologies to combat unauthorized automated data scraping. . . . The technological restrictions Meta uses to combat scraping and 'suspicious' activity include its lockout mechanism . . . and deploying machine-learning models and tools that detect and block automated scraping and suspicious account activity including 'inauthentic behavior, compromised accounts, and automated accounts.'" (citations omitted)).

128. 31 F.4th 1180 (9th Cir. 2022).

technological measures used to prevent third-party access to the platform's data. LinkedIn, the court states, revised its robots.txt file to "prohibit access to LinkedIn servers via automated bots."[129] It also uses "several technological systems to detect suspicious activity and restrict automated scraping," including "LinkedIn's Quicksand system," which detects "non-human activity indicative of scraping."[130] Moreover, LinkedIn's "Sentinel system throttles (slows or limits) or even blocks activity from suspicious IP addresses; and its Org Block system generates a list of known 'bad' IP addresses serving as large-scale scrapers."[131] "In total," the court concluded, "LinkedIn blocks approximately 95 million automated attempts to scrape data every day . . . ."[132]

C. From Contractual Restrictions to Criminal Liability?

The extent to which unauthorized access to platform data could trigger criminal liability under the CFAA may bolster the deterring impact of platforms' contractual restrictions on researchers. In *Sandvig*, the court ruled that the plaintiffs' specific research plans do not violate the access provision of the CFAA, explaining that this law does not criminalize mere terms-of-use violations on consumer websites.[133]

Outside the context of academic research, the Supreme Court restricted the scope of the CFAA, holding that it does not criminalize access made for an improper purpose when access itself was not unauthorized.[134] In *Van Buren v. United States*, a police officer was convicted under the CFAA for looking up license plates for improper monetary purposes using a law enforcement database he was authorized to access.[135] The Court held that an individual "exceeds authorized access" when "he accesses a computer with authorization but then obtains information located in particular areas of the computer—such as files, folders, or databases—that are off limits to him."[136] Accordingly, since the police officer was authorized to use the database to retrieve license plate information, he didn't exceed authorized access.[137]

Following *Van Buren*, in *hiQ*, the Ninth Circuit found the CFAA inapplicable to the defendants' unauthorized use of the plaintiff's data because the data was accessible to the general public.[138] Accessing publicly available information, the court reasoned, cannot be "unauthorized,"[139] but accessing sites that are restricted to users who sign in to the platform with their username and password could be, implying that researchers nevertheless face serious threats of criminal liability for accessing and using platforms' data in violation of their contractual obligations. As clearly demonstrated by recent lawsuits filed against companies who rely on large-

129. *Id.* at 1186.
130. *Id.*
131. *Id.* (footnote omitted).
132. *Id.*
133. Sandvig v. Barr, 451 F. Supp. 3d 73 (D.D.C. 2020).
134. Van Buren v. United States, 593 U.S. 374, 1662 (2021).
135. *Id.* at 1652–53.
136. *Id.* at 1662.
137. *Id.*
138. hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1199–1200 (9th Cir. 2022).
139. *Id.*

scale scraping of platform data to train models of Generative AI, such as OpenAI, CFAA claims against unauthorized scraping are here to stay.[140] Yet, unlike highly profitable commercial companies like OpenAI, who are willing to and financially capable of fighting back, researchers, who often enjoy strict research budgets, would probably prefer to avoid any potential risks of suits.[141] Nevertheless, if researchers had a valid legal defense against allegations of ToS violations, and if they could legally mandate platforms to remove some of the technological barriers they place on their sites, society as a whole could better enjoy public benefit research. Hence, in the next Part, we turn to propose how private law doctrines may establish such opportunities.

III. RECLAIMING RESEARCHERS' RIGHT TO ACCESS PLATFORM DATA IN THE UNITED STATES

In recent years, researchers have demonstrated platform data's potential in infusing scientific research with fresh energy and providing scientists with tools to develop novel research methodologies, perspectives, insights, and understandings. Nevertheless, platforms systematically use contractual and technological mechanisms to lock the data they host on their computers and deny access to it, even when access is intended solely for noncommercial, noncompeting, and socially beneficial scientific purposes.

In this Part, we argue that platforms' practice of "locking" data and indiscriminately excluding others from it stands in opposition to the letter and the spirit of private law. Section A advances this claim on the basis of the Digital Millennium Copyright Act of 1998 ("Copyright Act"). It shows that, for most types of platform data, platforms' attempt to use their ToS to exclude others should be preempted, and access and use of platform data for scientific purposes should be permitted under the fair use doctrine.

In Section B, we turn to the common law perspective, showing that it too reaches similar conclusions to those presented in Section A. In particular, we argue that contractual terms limiting access to platform data irrespective of its purpose can and should be narrowly interpreted by the courts, making them inapplicable to data access for scientific purposes. We further suggest that courts build on existing

---

140. See, for example, the class action complaint filed against OpenAI: Complaint, P.M. v. OpenAI LP, No. 3:23-cv-03199 (filed June 28, 2023).

141. Scientists, especially when employed by a research institution, are also rarely paid for conducting specific research or using a particular methodology. *See* Bruns, *supra* note 8, at 1557 ("[R]esearchers frustrated with the diminishing functionality of platform APIs, the disingenuous attacks on their work, and the ill-constructed alternative data lotteries orchestrated by the platforms and affiliated entities may be tempted simply to walk away."). From a rational choice (economic) perspective, then, the use of platform data exposes researchers to the risk of substantial personal (as well as institutional) liability but offers little if any material gains. At most, the products of such research may increase researchers' reputation, thereby advancing their careers (and earning) prospects, but a similar outcome would often be possible by other means that do not expose the researcher to similar risks.

nuisance and anti-enclosure jurisprudence to oblige the removal of technological barriers erected by platforms to prevent researchers from accessing platform data.

Properly applied, therefore, private law could offer researchers a meaningful shield against civil liability when accessing platform data for scientific purposes. However, it does not provide them with an affirmative right to access such data. Private law is often (though not always) hesitant to impose affirmative rights and duties among strangers, which platforms and scientists essentially are. Therefore, despite its desirability, we consider courts unlikely to impose an affirmative duty on platforms to provide researchers with the data they seek. For such a duty to exist, we argue that legislative efforts are typically necessary. Indeed, the EU attempted to create such a duty in the recently enacted Digital Service Act. Accordingly, Section C explores the EU's and similar regulatory frameworks proposed to create an affirmative right for scientists to access platform data and discuss their advantages and shortcomings.

A. Copyright Preemption

Copyright law may offer a critical legal basis for researchers to strike down breach-of-contract claims raised against them by platforms based on their ToS. The digital transformation has enabled new approaches to governing the use of copyrighted materials. To protect their rights and interests in the digital era of fast and easy piracy, rights holders turned to digital technology to lock their copyrighted works.[142] Digital locks, such as encryption and password-protected paywalls, backed by the anti-circumvention rules established under the Copyright Act,[143] restricted users' access to copyrighted works. These anti-circumvention rules impose civil and sometimes even criminal liability for circumventing these locks.[144] Digital access also facilitated an efficient contract formation between copyright rights holders and licensees, which allows copyright holders to contractually restrict certain uses of copyrighted materials, sometimes in contradiction to copyright norms.[145] The enforceability of such private ordering practices has been a subject of long-standing debate among legal scholars.[146] Scholars have raised concerns about the potential

142. *See, e.g.*, Viktor Mayer-Schönberger, *Beyond Copyright: Managing Information Rights with DRM*, 84 DENV. U. L. REV. 181 (2006). Rights holders used Digital Right Management to fight back against the misuse of duplication technologies used to copy copyrighted content easily and without authorization and peer-to-peer software used to share copyrighted information without rights holders' consent. *Id.* at 181–82.

143. *See generally* David Nimmer, *A Riff on Fair Use in the Digital Millennium Copyright Act*, 148 U. PA. L. REV. 673 (2000). Under 17 U.S.C. § 1201, a person who circumvents or traffics in products meant to circumvent an access control measure used to protect a copyrighted work will be in violation of the Copyright Act.

144. 17 U.S.C. § 1203 sets civil remedies for a violation of the anti-circumvention provisions including temporary or permanent injunctions and either actual or statutory damages. Treble damages may also be awarded for repeat violations. *Id.* § 1203(c)(4). Additionally, any person willfully violating the anti-circumvention provisions for commercial advantage or private gain is open to criminal liability under 17 U.S.C. § 1204.

145. *See* Amit Elazari Bar On, *Unconscionability 2.0 and the IP Boilerplate*, 34 BERKELEY TECH. L.J. 567, 595–612 (2019).

146. *See*, *e.g.*, David Nimmer, Elliot Brown & Gary N. Frischling, *The Metamorphosis of*

chilling effect of such restrictive actions on the fundamental goals of copyright law, which are aimed at promoting progress.[147] They have also expressed worries that contractual terms that provide broader protection to right holders beyond what is secured by copyright law may compromise socially beneficial practices such as tinkering,[148] testing, criticism, research, and learning,[149] and thus further undermine the goals of copyright law.

As we explain below, preempting contractual allegations that effectively override copyright law may transform the legal dispute between platforms and researchers from a contractual cause of action subject to state law into a federal copyright dispute. Under copyright law, scraping data for research purposes either involves legitimate reproduction of unprotected facts or is otherwise permissible under the fair use doctrine. The copyright preemption doctrine may, thus, facilitate platform-based research which could otherwise be contractually banned.[150]

1. The Copyright Preemption Doctrine

Section 301(a) of United States Code Title 17 states that copyright protection arises exclusively under federal law, preempting state law claims that involve rights that are "equivalent to any of the exclusive rights within the general scope of copyright" and that "come within the subject matter of copyright."[151] This preemption doctrine effectively limits the freedom of the parties to contract around copyright law in order to secure for themselves one (or more) of the exclusive rights granted by law to copyright holders.[152]

Contract into Expand, 87 CALIF. L. REV. 17, 20 (1999). *But see* Joel Rothstein Wolfson, *Contract and Copyright Are Not at War: A Reply to "the Metamorphosis of Contract into Expand"*, 87 CALIF. L. REV. 79 (1999).

147. *See, e.g.*, Niva Elkin-Koren, *Can Formalities Save the Public Domain? Reconsidering Formalities for the 2010s*, 28 BERKELEY TECH. L.J. 1537, 1537–38 (2013); Margaret Jane Radin, *Regime Change in Intellectual Property: Superseding the Law of the State with the "Law" of the Firm*, 1 U. OTTAWA L. & TECH J. 173, 178 (2004); Niva Elkin-Koren, *A Public-Regarding Approach to Contracting Copyrights*, *in* EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY FOR THE KNOWLEDGE SOCIETY 191, 192 (Rochelle Cooper Dreyfuss, Diane Leenheer Zimmerman & Harry First eds., 2001) [hereinafter Elkin-Koren, *A Public-Regarding Approach*].

148. Pamela Samuelson, *Freedom to Tinker*, 17 THEORETICAL INQUIRIES L. 563, 564 (2016); Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 198–202 (2017).

149. *See generally* Lemley & Casey, *supra* note 20.

150. *See generally* 17 U.S.C. § 301; Mark A. Lemley, *Beyond Preemption: The Law and Policy of Intellectual Property Licensing*, 87 CALIF. L. REV. 111 (1999).

151. 17 U.S.C. § 301(a) ("[A]ll legal or equitable rights that are equivalent to any of the exclusive rights within the general scope of copyright as specified by section 106 in works of authorship that are fixed in a tangible medium of expression and come within the subject matter of copyright as specified by sections 102 and 103 . . . are governed exclusively by Title 17.").

152. Niva Elkin-Koren, *Copyright in the Digital Ecosystem: A User Rights Approach*, *in* COPYRIGHT LAW IN AN AGE OF LIMITATIONS AND EXCEPTIONS 132, 167 (Ruth Okediji ed., 2017).

To determine whether a specific contractual claim should be preempted, courts apply a two-part analysis[153]: First, the court examines whether the work that would be affected by the plaintiff's exercise of a state-created right comes within the subject matter of copyright, as specified by sections 102 and 103 of the Copyright Act.[154] If the subject matter of the state law claim falls within copyright subject matter, then the court must move to examine the second test—to determine whether the rights asserted under the state law claim are equivalent to any of the exclusive rights listed under section 106 of the Copyright Act.[155]

To meet the first prong of the analysis—the "subject matter requirement"—the underlying work must be "a 'literary work[ ],' a 'musical work[ ],' a 'sound recording[ ],' or any other category of 'work[ ] of authorship' within the 'subject matter of copyright.'"[156] Importantly, the scope of copyright for the purpose of preemption is broader than the scope of available copyright protection. Hence, a plaintiff may not claim a contractual right in a work of a *type* covered by sections 102 and 103 of the Copyright Act, even if for some reason that work is not protected by copyright (i.e., because it fell into the public domain or because it lacks sufficient originality).[157] Such a broad interpretation procures the statutory distinction between copyrighted works and works that ought to remain in the public domain, limiting the ability of the parties to contract around it.

To meet the second prong of the preemption test—the "equivalence" requirement—the defendant must show that the right the plaintiff asserts in the work (which meets the first prong of the test) is "equivalent to any of the exclusive rights within the general scope of copyright as specified by section 106."[158] For preemption to apply, "the state law claim must involve acts of reproduction, adaptation, performance, distribution or display."[159] Nevertheless, a claim is not preempted if it "include[s] any extra elements that make it qualitatively different from a copyright infringement claim."[160] To determine whether such an extra element exists, courts evaluate "what the plaintiff seeks to protect, the theories in which the matter is thought to be protected and the rights sought to be enforced."[161] This requires "a holistic evaluation of the nature of the 'rights sought to be enforced,' and a determination whether the state law action 'is qualitatively different from a copyright infringement claim.'"[162]

---

153. ML Genius Holdings, LLC v. Google LLC, No. 20-3113, 2022 WL 710744, at *2 (2d Cir. Mar. 10, 2022).

154. *See In re* Jackson v. Roberts, 972 F.3d 25, 42 (2d Cir. 2020).

155. *Genius*, 2022 WL 710744, at *3.

156. *In re Jackson*, 972 F.3d at 42 (alterations in original) (citation omitted).

157. Forest Park Pictures v. Universal Television Network, Inc., 683 F.3d 424, 429–30 (2d Cir. 2012).

158. *In re Jackson*, 972 F.3d, at 43 (emphasis omitted) (quoting 17 U.S.C. § 301(a)).

159. *Genius*, 2022 WL 710744, at *3 (quoting Briarpatch Ltd., L.P. v. Phoenix Pictures, Inc., 373 F.3d 296, 305 (2d Cir. 2004)).

160. *Id.* (alteration in original) (quoting *Briarpatch*, 373 F.3d at 305).

161. *Id.* at *7 (quoting *Briarpatch*, 373 F.3d at 306).

162. *In re Jackson*, 972 F.3d at 44 n.17 (emphasis omitted) (quoting Comput. Assocs. Int'l, Inc. v. Altai, 982 F.3d 693, 716 (2d Cir. 1992)).

How does platforms' contractual right to control access to and use of data on their servers fit within this two-prong test? As we demonstrate below, this analysis depends on the type of platform data being scraped.

2. User Produced Data (UPD), Digital Footprints, and Preemption Analysis

As we explained earlier, one type of platform data is UPD—that is, data that originates in third parties, such as users' Facebook posts, ads, photos, videos, and similar content.[163] Also included in this category is information about users' preferences, locations, behavior, and related information. Such information is reflected in users' online activity on the platform, which leaves valuable digital footprints behind it.[164] Many research studies engage in scraping UPD[165] and collecting and analyzing users' digital footprints,[166] which have also become a critical input for training ML models.[167] Would researchers who scrape UPD and users' digital footprints succeed in preempting platforms' breach of contract claim?

a. UPD and Preemption Analysis

UPD meets the first prong of the preemption test quite easily. UPD is copyrightable because it falls within the subject matter of copyright protection as literary work (e.g., posts), audiovisual work (e.g., videos), pictorial work, or graphic work (e.g., photos). That the copyright in this type of data is not the platforms'—but rather their users'[168]—should not be held against preemption. Recently, in *ML Genius Holdings LLC v. Google LLC*, the Second Circuit applied copyright preemption to a breach-of-contract claim raised by a plaintiff who did not own the copyright in the underlying work.[169] In that case, the plaintiff, ML Genius, owned a website that displayed lyric transcriptions of songs that were compiled by music fans

---

163. *See supra* Section I.B.

164. *See, e.g.*, Ben Lutkevich, *Digital Footprint*, TECHTARGET, https://www.techtarget.com/whatis/definition/digital-footprint [https://perma.cc/DYU2-RAR6] (Feb. 2023).

165. *See, e.g.*, Anna Ruelens, *Analyzing User-Generated Content Using Natural Language Processing: A Case Study of Public Satisfaction with Healthcare Systems*, 5 J. COMPUTATIONAL SOC. SCI. 731 (2022).

166. *See, e.g.*, Scott A. Golder & Michael W. Macy, *Digital Footprints: Opportunities and Challenges for Online Social Research*, 40 ANN. REV. SOCIO. 129 (2014).

167. *See, e.g.*, Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei & Ilya Sutskever, *Language Models Are Unsupervised Multitask Learners*, OPENAI (2019), https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf [https://perma.cc/8ZBK-KZD3] (describing the large amounts of training data, including UGD, that was needed to develop the ChatGPT).

168. *See, e.g.*, *Terms of Service Overview 3.3(2)*, META (Jan. 1, 2025), https://www.facebook.com/legal/terms/ [https://perma.cc/7RET-SLHB] ("[W]hen you share, post, or upload content that is covered by intellectual property rights on or in connection with our Products, you grant us a non-exclusive, transferable, sub-licensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content . . . .").

169. No. 20-3113, 2022 WL 710744, at *1–3 (2d Cir. Mar. 10, 2022).

or obtained directly from artists. Because the plaintiff did not hold the copyrights for the original lyrics, he paid the copyright holders for licenses to publicly display them on his website.[170] In fact, platforms' attempt to contractually ban the use of UPD, despite being only non-exclusive licensees of such content, may actually support preemption. As the court noted in *X Corp. v. Bright Data Ltd.*, this contract strategy undermines copyright law by interfering with the users' ability to exercise their exclusive rights as copyright owners.[171]

As to the second prong of the preemption test, the right platforms may assert in UPD applies to any reproduction of the data, including to any derivative modifications of the data made to make it suitable for training models.[172] Whether the fact that this right stems from contract law and not from copyright is sufficient as an "extra element" necessary to avoid preemption is subject to a major circuit split.[173]

Some courts take a stricter approach against preemption, practically upholding any contractual provision that applies to those parties who agreed to accept it. In *ProCD, Inc. v. Zeidenberg*, for instance, the court held that a breach-of-contract claim was not preempted because a "copyright is a right against the world," while "[c]ontracts, by contrast, generally affect only their parties" and therefore "do not create 'exclusive rights.'"[174] Under this line of reasoning, a breach-of-contract claim would survive unless the respondent may prove that no valid contract was formed. Such contract formation issues may arise, for instance, in "browsewrap" agreements, where questions of awareness to the contractual restrictions and explicit consent may require elaboration.[175] However, as a general matter, courts routinely enforce platforms' ToS as binding contracts.[176]

To the contrary, other courts favor preemption, requiring more than contractual privity to transform an otherwise equivalent claim into one that is qualitatively

---

170. Petition for Writ of Certiorari, *Genius*, 2022 WL 710744 (No. 22-121), 2022 WL 3227953, at *7–9.

171. 733 F. Supp. 3d 832, 845 (N.D. Cal. 2024).

172. *See, e.g.*, Nicola Lucchi, *ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems*, 15 EURO. J. RISK REG. 602, 618, 623 (2024).

173. Petition for Writ of Certiorari, *Genius*, 2022 WL 710744 (No. 22-121), 2022 WL 3227953, at *20–23.

174. 86 F.3d 1447, 1454 (7th Cir. 1996); *see also* Bowers v. Baystate Techs., Inc., 320 F.3d 1317, 1325 (Fed. Cir. 2003), *cert. denied*, Bystate Techs., Inc. v. Bowers, 539 U.S. 928 (2003); Lipscher v. LRP Publ'ns, Inc., 266 F.3d 1305, 1318 (11th Cir. 2001) (finding that "claims involving two-party contracts are not preempted because contracts do not create exclusive rights, but rather affect only their parties" and holding that a contract claim was not preempted (citations omitted)).

175. Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1176 (9th Cir. 2014) ("The defining feature of browsewrap agreements is that the user can continue to use the website or its services without visiting the page hosting the browsewrap agreement or even knowing that such a webpage exists." (internal quotations and citation omitted)).

176. *See* Hancock v. AT&T Co., 701 F.3d 1248, 1256 (10th Cir. 2012); Specht v. Netscape Commc'ns Corp., 306 F.3d 17, 32–35 (2d Cir. 2002); Serrano v. Cablevision Sys. Corp., 863 F. Supp. 2d 157, 164 (E.D.N.Y. 2012); *see generally* MARGARET J. RADIN, BOILERPLATE: THE FINE PRINT, VANISHING RIGHTS, AND THE RULE OF LAW (2013).

different from a copyright infringement claim.[177] The Second Circuit followed this view in the *ML Genius* case.[178] Genius argued that since its breach-of-contract claim requires it to plead "mutual assent and valid consideration," and "assert[] rights only against the contractual counterparty, not the public at large," its claim is qualitatively different from a copyright claim, and therefore should not be preempted.[179] The Second Circuit, however, disagreed, criticizing that Genius is effectively contemplating "a per se rule that all breach of contract claims are exempt from preemption." Such a rule, according to the court, "would be in tension with our precedent holding that the general scope inquiry is 'holistic.'"[180] The court concluded that Genius effectively sought to protect a right that is coextensive with the exclusive right to reproduce and make derivative works of the lyrics, and therefore it is preempted.[181]

Similarly, a contractual provision in platforms' ToS that restricts the scraping of UPD functions like a "right against the world."[182] While technically it applies only to people who accept it, all platforms' *users* are bound by it. That is, if you use the platform, you are bound by its ToS. Additionally, there is often no alternative ways to obtain this data but through the platform.[183] For instance, whoever seeks to access a post on Facebook must do it through Facebook's infrastructure, which is accessible only to signed-up users.[184] Moreover, as the court noted in *X Corp. v. Bright Data*, platforms' ToS do not reflect "an arm's length contract between two sophisticated parties in which one or the other adjusts their rights and privileges under federal copyright law," but rather "a massive regime of adhesive terms imposed by X Corp. that stands to fundamentally alter the rights and privileges of the world at large (or at least hundreds of millions of alleged X users)."[185] Thus, in restricting users' ability

---

177. *See, e.g.*, Universal Instruments Corp. v. Micro Sys. Eng'g, Inc., 924 F.3d 32, 49 (2d Cir. 2019) (explaining that the parties' "contractual privity does nothing to change the fact that vindication of an exclusive right under the Copyright Act" asserted through a breach-of-contract claim "is preempted by the Copyright Act"); Wrench LLC v. Taco Bell Corp., 256 F.3d 446, 457 (6th Cir. 2001) ("If the promise amounts only to a promise to refrain from reproducing, performing, distributing or displaying the work, then the contract claim is preempted.").

178. ML Genius Holdings LLC v. Google LLC, No. 20-3113, 2022 WL 710744, at *3 (2d Cir. Mar. 10, 2022).

179. *Id.* at *4 (alteration in original) (internal quotations and citations omitted).

180. *Id.* (citations omitted).

181. *Id.*

182. Elkin-Koren, *supra* note 65, at 103.

183. *Id.* at 104 ("[T]he introduction of new distribution technologies blurs the distinction between rights in personam and rights in rem. The availability of direct communication with users and the technical ability to prevent any unlicensed access by technological fencing facilitate a regime that is very similar in its nature to a property regime."). *But see* Meta Platforms, Inc. v. Bright Data Ltd., No. 23-cv-00077-EMC, 2024 WL 251406 (N.D. Cal. Jan. 23, 2024). Construing Meta's ToS, the court concluded that the contractual terms restricting scraping do not apply to Bright Data's logged-off scraping of publicly viewable data, and therefore denied Meta's motion for summary judgment. *Id.*

184. *Terms of Service Overview 3.1*, META, (Jan. 1, 2025), https://www.facebook.com/legal/terms/ [https://perma.cc/8V7C-7U9Y].

185. 733 F. Supp. 3d 832, 850 (N.D. Cal. 2024).

to reproduce and make derivations of copyrightable material on their servers, platforms *de facto* secure to themselves a proprietary right in the material. Their breach-of-contract claims should therefore be preempted. The dispute between the platform and the researchers would, thus, be governed by copyright law, under which reproduction of copyrighted material for research purposes may be permissible pursuant to the fair use doctrine.[186]

b. Users' Data Trail and Preemption Analysis

Applying copyright preemption to users' data trails (such as users' profile data) might be more challenging because such raw data is not copyrightable. The Copyright Act explicitly excludes in section 102(b) ideas, procedures, processes, systems, methods of operation, concepts, principles, or discoveries from copyright protection.[187] Likewise, knowledge, truths ascertained, conceptions and ideas "are free as the air to common use."[188] Article 9.2 of the TRIPS Agreement also provides that "[c]opyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such."[189] As famously elaborated by the Court in *Feist Publications, Inc. v. Rural Telephone Service*, the idea-expression dichotomy in copyright law ensures that only original expressions of facts—but not facts in themselves—are protected by copyright.[190] These statutory exceptions reflect a delicate balance "between the interests of authors . . . in the control and exploitation of their writings and discoveries on the one hand, and society's competing interest in the free flow of ideas, information, and commerce on the other hand."[191]

Users' data trails arguably equate with facts. They essentially reflect an objective representation of facts about relevant users—what they like, where and when they go, who they follow, and other relevant details.[192] Therefore, from a copyright perspective, such raw (or unprocessed) data about users should remain free for all, and especially for researchers seeking to exploit it for public benefit research. But, if users' data trails are not copyrightable, how do they come "within the subject matter

---

186. 17 U.S.C. § 107 ("[T]he fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.").

187. 17 U.S.C. § 102(b).

188. Int'l News Serv. v. Associated Press, 248 U.S. 215, 250 (1918) (Brandeis, J., dissenting).

189. Agreement on Trade-Related Aspects of Intellectual Property Rights, art 9.2, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

190. 499 U.S. 340, 348 (1991) (noting that "all facts—scientific, historical, biographical, and news of the day . . . 'may not be copyrighted and are part of the public domain available to every person'" (quoting Miller v. Universal City Studios, Inc., 650 F.2d 1365, 1369 (5th Cir. 1981))).

191. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 429 (1984).

192. *See* Tesh W. Dagne, *Where Copyright Meets Privacy in the Big Data Era: Access to and Control over User Data in Agriculture and the Role of Copyright*, 24 VAND. J. ENT. & TECH. L. 675, 720–22 (2022).

of copyright," as required by the first prong of the preemption analysis?[193] We claim that users' data trails are subject to copyright law, which mandates their *exclusion* from copyright protection. Otherwise, if data was not copyrightable, it would have made little sense to explicitly exclude it from copyright law's protection. Copyright law applies to intangible creations, distinguishing between those that are sufficiently original and hence subject to exclusivity and those that ought to remain in the public domain as building blocks for future creators. When platforms attempt to assert proprietary rights in users' data trails, they essentially interfere with the delicate balance between exclusivity and access articulated by copyright policies.[194] A robust contractual ban on accessing and using users' data trails essentially allows platforms to attain property rights that cannot be owned by anyone under copyright law. Accordingly, users' data trails are subject to copyright because they are predominantly uncopyrightable and thus meet the first prong of the preemption test.

As to the second prong of the preemption doctrine, a contractual restriction on scraping (all types of) platform data, including users' data trails, is equivalent in its effect to possessing an exclusive right in such data, especially given that this data may not be replicated elsewhere. As noted, such a restriction embedded in boilerplate ToS binds all platforms' users and, therefore, it resembles the *in rem* characteristics of a property right.[195] Thus, it should be governed by copyright law, rather than by state contract law. Researchers who access and use platform users' data trails for scientific purposes should hence succeed in preempting contractual claims asserted against them.

A similar outcome could be reached by applying the doctrine of conflict preemption. Pursuant to this doctrine, "state laws that conflict with federal law are 'without effect,'"[196] and may not be applied in "those instances where the challenged state law 'stands as an obstacle to the accomplishment and execution of the full purposes and objectives of Congress.'"[197] As Guy Rub contemplates, copyright preemption fits within this notion of conflict preemption.[198] This notion was supported by the court in *X Corp. v. Bright Data*:

> Although conflict preemption has played second fiddle to express preemption in the caselaw as of late, it is the more appropriate consideration when the question presented is not whether rights created by state law are equivalent to rights created by federal copyright law but whether enforcement of state law undermines federal copyright law.[199]

---

193. *See supra* note 151 and accompanying text.

194. *See* Robert A. Kreiss, *Accessibility and Commercialization in Copyright Theory*, 43 UCLA L. REV. 1, 2–4 (1995).

195. *See supra* Section I.C.

196. Altria Grp., Inc. v. Good, 555 U.S. 70, 76 (2008) (quoting Maryland v. Louisiana, 481 U.S. 725, 746 (1981)).

197. Arizona v. United States, 567 U.S. 387, 399 (2012) (quoting Hines v. Davidowitz, 312 U.S. 52, 67 (2012)).

198. Guy A. Rub, *A Less-Formalistic Copyright Preemption*, 24 J. INTELL. PROP. L. 327, 329–31 (2017).

199. 733 F. Supp. 3d 832, 850 (N.D. Cal. 2024) (emphasis omitted).

Accordingly, since allowing platforms to robustly exclude researchers from scraping all platform data constitutes a critical obstacle to achieving the goals of copyright policy in promoting scientific research for public benefit, the contractual limitation on data scraping in platform ToS should be without effect.

To determine whether to sustain a breach-of-contract claim at the expense of achieving copyright policy, the conflict preemption scheme would (1) consider the specific considerations of the platform in setting the relevant contractual provision; and (2) examine how these considerations fit within the broader framework of copyright policy.[200] First, the key motive of platforms when deploying boilerplate ToS against data scraping is financial.[201] Thus, it is extremely close (and even identical) to the economic interests of copyright holders. Yet, even if we accept that such restrictions are also put in place to protect the privacy interests of platforms' users,[202] these interests may be adequately protected through alternative ethical mechanisms, and, therefore, their protection does not depend solely on contract law. Second, allowing platforms to acquire an unlimited, property-like right in data conflicts with the nature and operation of federal copyright law. Copyright monopoly is contingent, instrumental, and limited to the level necessary to provide sufficient incentives.[203] As we noted, copyright is restricted under statutory provisions and legal doctrines that exclude information from its protection.[204] It is further limited by the fair use doctrine, which permits unauthorized yet socially beneficial uses of protected works, such as research.[205] Platforms' attempt to strictly restrict researchers from reproducing raw data would allow them to control areas outside their monopoly.[206]

Before we end this Part, it is important to stress that oftentimes, as a technical matter, researchers would need to intermediately scrape full, copyright-protected, platform webpages to extract the non-copyrightable data trails of users. However, this should not expose them to liability for copyright infringement. More than a decade ago, in a different commercial context, a lower court found similar reproduction to be copyright infringement. In *Facebook, Inc., v. Power Ventures, Inc.*, the defendant Power.com provided a service that aggregated and displayed

---

200. *See* Guy A. Rub, *Moving from Express Preemption to Conflict Preemption in Scrutinizing Contracts over Copyrighted Goods*, 56 AKRON L. REV. 303, 315–17 (2023).

201. *See supra* note 87 and accompanying text.

202. *See supra* note 88 and accompanying text.

203. Elkin-Koren, *supra* note 65, at 100.

204. *X Corp.*, 733 F. Supp. 3d at 852. The court explained that upholding X's contractual claims against Bright Data's scraping of publicly available data "would give itself de facto copyright ownership over content that Congress declined to extend copyright protection to in the first place (e.g., likes, user names, short comments)," and it further described how this "shrinks the public domain, restricting free reproduction, adaptation, distribution, and display of publicly available, non-expressive material." *Id.*

205. *Id.* at 852. The court explained that "[o]nly by receiving permission and paying X Corp. could Bright Data, its customers, and other X users freely reproduce, adapt, distribute, and display what might (or might not) be available for taking and selling as fair use," while frustrating the operation of the fair use doctrine. *Id.*

206. Assessment Techs. of WI, LLC v. WIREdata, 350 F.3d 640, 645–48 (7th Cir. 2003) (striking down an attempt to restrict access to data that was not copyrighted).

social networking and email accounts on a single portal.[207] Technically, users of different social media platforms gave Power.com their login credentials, allowing Power.com's software to access their personal accounts and scrape data from the platforms' websites.[208] Facebook alleged that Power.com violated its copyright in scraping its websites, even though the latter merely sought to scrape users' data (such as their different social connections on various platforms).[209] The district court ruled in Facebook's favor, concluding that "if Defendants first have to make a copy of a user's entire Facebook profile page in order to collect that user content, such action may violate Facebook's proprietary rights."[210]

Yet researchers only copy with the purpose of pursuing public benefit research objectives—not to promote private commercial interests.[211] This is one important reason justifying a different outcome. Additionally, as we explained earlier, users' data trails are not subject to copyright protection because it is a raw material that ought to remain in the public domain. Generally, copyright law refuses to allow copyright holders to control areas outside their monopoly,[212] and any attempt to do so may be treated as copyright misuse.[213] In *Assessment Technologies v. Wiredata*, for instance, the Seventh Circuit refused to allow the plaintiff, AT, to claim copyright protection in the compiled data that was in the public domain. AT essentially created a database for tax authorities, which included information such as the age of the property and the number of rooms.[214] While the court found that AT's computer program (called "Market Drive") was sufficiently original to obtain copyright protection, it held that Wiredata didn't infringe copyright because it sought to obtain the raw data about the property to arrange it in a different way that will be useful for estate brokers.[215] The court explained that "the process of extracting the raw data from the database does not involve copying Market Drive, or creating, as AT mysteriously asserts, a derivative work; all that is sought is raw data, data created not by AT but by the assessors, data that are in the public domain."[216]

The fair use doctrine should shield researchers who engage in peripheral reproduction from copyright liability due to the transformative nature of their use, and because it does not threaten the platforms' core market.[217] Researchers essentially engage in "non-expressive" copying.[218] They do not intend to compete

207. No. C 08–5780 JF (RS), 2009 WL 1299698, at *1–2 (N.D. Cal. May 11, 2009).
208. *Id.*
209. *Id.*
210. *Id.* at *4.
211. *Id.*
212. 17 U.S.C. § 103(b) ("The copyright in a compilation or derivative work extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work, and does not imply any exclusive right in the preexisting material."); *see also* Assessment Techs. of WI, LLC v. WIREdata, 350 F.3d 640, 645–48 (7th Cir. 2003) (striking down an attempt to restrict access to data that was not copyrighted).
213. *Assessment Techs. of WI, LLC*, 350 F.3d at 645.
214. *Id.* at 642.
215. *Id.* at 643–44.
216. *Id.* at 644.
217. Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 6 J. COPYRIGHT SOC'Y U.S.A. 291, 303–10 (2019).
218. Lemley & Casey, *supra* note 20, at 760 n.92, 778 n.189.

with the platform in the data market, nor do they seek to develop competing products or services. If researchers were required to obtain a license for non-expressive copying of platforms' webpages, they probably wouldn't (for lack of financial resources), and we couldn't enjoy the fruits of their studies.[219] Thus, even if researchers engage in unauthorized copying in the course of obtaining access to non-protected data, they are immune from copyright infringement claims under the fair use doctrine.

3. Co-Produced Data (CPD), Platform Produced Data (PPD), and Preemption Analysis

A different type of platform data is PPD, which may include processed data or data aggregations. For instance, the arrangement of information about rental units and their availability on Airbnb[220] and the set of insights metrics aggregated across all instant articles that have been published by a Facebook page are examples of PPD.[221] How does preemption analysis apply to such data?

As to the first prong of the preemption test, data aggregations are a sort of *compilations* and hence of the *type* protected by copyright[222] to the extent that they reflect original choices in the selection or arrangement of their underlining components.[223] To obtain copyright protection, these choices must be sufficiently "original"—a minimal requirement that enables a work to be distinguished from similar works that are in the public domain.[224] However, even if the degree of originality reflected in these choices is insufficient for the purpose of copyright protection, it would nonetheless suffice to meet the first prong of preemption.[225] Indeed, "[a]s long as a work fits within one of the general subject matter categories of sections 102 and 103, the bill prevents the States from protecting it even if it fails

---

219. *Id.* at 770 n.150.

220. *See, e.g.*, Teresa Scassa, *Ownership and Control over Publicly Accessible Platform Data*, 43 ONLINE INFO. REV. 986 (2019). Scassa explains that the data hosted on the Airbnb site can be scraped and analyzed so as to provide important insights into many issues, such as "the platform's effects on the cost and availability of long-term accommodation, its impact on incumbent short-term accommodation providers, the incidence of discrimination in Airbnb rentals and pricing and the extent to which the platform is used to support full scale commercial ventures." *Id.* at 986.

221. *Insights Metrics of Aggregated Instant Articles*, META (Nov. 30, 2015), https://developers.facebook.com/docs/graph-api/reference/v22.0/instant-articles-insights-aggregated [https://perma.cc/3QTY-EYJU].

222. 17 U.S.C. § 103.

223. *See* Feist Publ'ns, Inc. v. Rural Teleph. Serv. Co., 499 U.S. 340, 348–49; John F. Hayden, *Copyright Protection of Computer Databases After* Feist, 5 HARV. J.L. & TECH. 215, 225 (1991).

224. Bucklew v. Hawkins, Ash. Baptie & Co., 329 F.3d 923, 929 (7th Cir. 2003).

225. Durham Indus., Inc. v. Tomy Corp., 630 F.2d 905, 919 n.15 (2d Cir. 1980) (holding that § 301 of the Copyright Act "prevents the States from protecting [a work] even if it fails to achieve Federal statutory copyright because it is too minimal or lacking in originality to qualify" (internal quotation and citation omitted)).

to achieve Federal statutory copyright because it is too minimal or lacking in originality to qualify."[226]

As to the second prong of preemption, platforms' robust contractual bans on accessing and using PGD is equivalent in its effect to an exclusive right in such data, which may not be obtained but through the platform.[227] Therefore, such contractual claims asserted against researchers should be subject to copyright preemption. Researchers, as a result, would have a good shield against contractual claims asserted against them for using PPD in violation of platforms' ToS. But is this enough to shield them from platforms' proprietary allegations?

Platforms, as noted earlier, may raise a claim of copyright infringement against researchers who use PPD without authorization.[228] Yet even if researchers do extract protected expressions in the course of using PPD, the doctrine of fair use should suffice to defend them.[229]

More challenging, however, would be potential allegations of misappropriation under trade secret law. Generally, to raise a trade secret cause of action, it is necessary to prove: (1) the existence of a trade secret—secret information from which the owner derives independent economic value and which the owner takes reasonable steps to keep secret; and (2) misappropriation of the trade secret by improper means, such as theft, bribery, misrepresentation, and espionage.[230] It is questionable if publicly accessible PPD satisfies the secrecy requirement.[231] However, such information may be subject to password login requirements or licenses backed up by ToS, which may limit their accessibility.[232]

The case of *Compulife Software, Inc. v. Newman*, for instance, concerned Compulife's software that generates life insurance quotes by relying on Compulife's factual compilation (database) of insurance rates.[233] Compulife alleged that Newman supervised a scraping attack of its database to get many millions of quotes—far more than a human could ever physically obtain.[234] The Eleventh Circuit recognized the secrecy of Compulife's database as a whole because accessing it was conditioned upon holding a license.[235] Moreover, the court noted that "even if individual quotes

---

226. H.R. REP. NO. 94-1476, at 131 (1976).

227. Elkin-Koren, *supra* note 65, at 104.

228. *See supra* Section II.C.

229. *See supra* note 186 and accompanying text.

230. *See* 18 U.S.C. § 1839(3), (5), (6); UNIFORM TRADE SECRET ACT § 1(2), (4) (NAT'L CONF. OF COMM'RS ON UNIF. STATE L.).

231. Peter J. Toren, *A Dubious Decision: Eleventh Circuit Finds Scraping of Data from a Public Website Can Constitute Theft of Trade Secrets (Part I)*, IPWATCHDOG (July 2, 2020, 4:15 PM), https://www.ipwatchdog.com/2020/07/02/dubious-decision-eleventh-circuit-finds-scraping-data-public-website-can-constitute-theft-trade-secrets-part/id=123029/ [https://perma.cc/JD5V-PGXR].

232. Geoffrey Xiao, *Data Misappropriation: A Trade Secret Cause of Action for Data Scraping and a New Paradigm for Database Protection*, COLUM. SCI. & TECH. L. REV. 125, 145–48 (2022).

233. 111 F.4th 1147, 1153 (11th Cir. 2024).

234. *Id.* at 1155.

235. *Id.* at 1161. ("So long as the precautions taken were reasonable, it doesn't matter that the defendant found a way to circumvent them. Indeed, even if the trade-secret owner took no measures to protect its secret from a certain type of reconnaissance, that method may still

that are publicly available lack trade secret status, the whole compilation of them (which would be nearly impossible for a human to obtain through the website without scraping) can still be a trade secret."[236] Whether other courts will adopt a similar broad interpretation of the secrecy requirement is currently unknown. The court further found that the nature of the data scraping used by the defendants to reproduce the protected database constitutes unlawful misappropriation.[237] While the court noted that "scraping and related technologies (like crawling) may be perfectly legitimate,"[238] the scraping methods used by the defendants were not.[239] Likewise, scraping executed in violation of platforms' ToS may also constitute "improper means." In that case, misappropriation of PPD could constitute a barrier to platform data research, unless courts would give special weight to the fact that researchers scrape the data for noncommercial, public beneficial purposes.[240]

To conclude this Section, platforms' attempt to secure to themselves proprietary rights in platform data by imposing robust contractual restrictions on accessing and using the data on their computers should fail when access is sought for noncommercial, public benefit research. Copyright preemption should defend researchers from contractual causes of actions because platform data falls within the subject of copyright—either as copyright-protected users' content or copyright-protected data aggregations—or because copyright deliberately excludes it from copyright protection. Other legal claims potentially raised by platforms—based on copyright infringement or misappropriation—should also fail, considering the fair and socially beneficial context of platform data research.

B. The Common Law of Access to Data: From Contracts to Nuisance

The preemption doctrine offers scientists using platform data a property-based defense against prospective breach-of-contract claims by digital platforms. In this Section, we turn to contract law to show that the proper interpretation of platforms' ToS, as well as the unconscionability doctrine, support a similar conclusion. Private law, as Subsection 3 further shows, also provides researchers with claims to require the removal of technological barriers to platform data.

---

constitute improper means." (quoting Compulife Software Inc. v. Newman, 959 F.3d. 1288, 1312 (11th Cir. 2020))).

236. *Id.* at 1162 (citing *Compulife*, 959 F.3d at 1314).

237. *Id.* at 1163. The court found that the defendant "copied the order of Compulife's copyrighted code and used that code to commit a scraping attack that acquired millions of variable-dependent insurance quotes" and that "this deceptive behavior resembles the acquisition of a trade secret through surreptitious aerial photography." *Id.* (emphasis omitted).

238. *Id.* at 1162 (emphasis omitted).

239. *Id.* at 1163 ("[T]he defendants in this case did not take innocent screenshots of a publicly available site; instead, they copied the order of Compulife's copyrighted code and used that code to commit a scraping attack that acquired millions of variable-dependent insurance quotes. If they had not formatted and ordered their code exactly as Compulife did, they would not have been able to get the millions of quotes that they got.").

240. RESTATEMENT (THIRD) OF UNFAIR COMPETITION § 43 cmt. c. (AM. L. INST. 1995) ("The propriety of the acquisition must be evaluated in light of all the circumstances of the case, including whether the means of acquisition are inconsistent with accepted principles of public policy . . . .").

1. Interpreting Clauses Locking-in Platform Data

Part III showed that digital platforms use their ToS to place contractual barriers and lock in platform data. Typically, clauses restricting access to data are drafted in broad terms, failing to distinguish between the different entities and purposes for which access is sought. TikTok, for example, prohibits *all* users from making "unauthorised copies," and users may not "modify, adapt, translate, reverse engineer, disassemble, decompile or create any derivative works of the Services or any content included therein";[241] and Meta states that "[y]ou may not access or collect data from our Products using automated means,"[242] irrespective of who collects the data and for what exact purpose.

Contractual restrictions on access to data are often drafted in absolute terms. Yet a basic insight of modern contract law is that a promise to do (or refrain from doing) something is not necessarily a promise to perform without exception.[243] Instead, contract law may employ various doctrines and principles to limit the scope of contractual obligations, such as the doctrines of impossibility, impracticability, and frustration of purpose,[244] and the duty of good faith in performance.[245]

Interpretation is another way contract law may limit contractual obligations, even when no such limitation appears in the written text. Applied to boilerplate agreements, such as ToS, contract law provides researchers with multiple interpretive rules to defend against breach-of-contract claims, such as interpretation against the drafter[246] and interpretation reflecting the parties' reasonable intentions and expectations.[247] For example, though many ToS grant platforms a seemingly unrestricted power to unilaterally modify the terms of the agreement, courts have narrowly interpreted such clauses, finding modifications unenforceable without prior notice, even when no notice requirement appeared in the contractual text.[248]

---

241. *See Terms of Service 5*, TIKTOK, https://www.tiktok.com/legal/page/us/terms-of-service/en [https://perma.cc/2QS4-DHH7] (Nov. 2023).

242. *See Terms of Service 3.23*, FACEBOOK, (Jan. 1, 2025), https://www.facebook.com/terms.php [https://perma.cc/Q8YH-US5N].

243. CHARLES FRIED, CONTRACT AS PROMISE: A THEORY OF CONTRACTUAL OBLIGATION 9–10 (1981).

244. *See* RESTATEMENT (SECOND) OF CONTRACTS §§ 152, 261–72 (AM. L. INST. 1981).

245. *Id.* at § 205; Market Street Assocs. L.P. v. Frey, 941 F.2d 588, 592–96 (7th Cir. 1991).

246. RESTATEMENT (SECOND) OF CONTRACTS § 206 (AM. L. INST. 1981); RESTATEMENT OF THE LAW, CONSUMER CONTRACTS § 4 (AM. L. INST. 2024).

247. *See* Sutton v. East River Sav. Bank, 55 N.Y.S.2d 550, 555 (1982) ("Our goal must be to accord the words of the contract their 'fair and reasonable meaning.'" (citation omitted)); Vector Cap. Corp. v. Ness Techs., Inc., No. II Civ. 6259(PKC), 2012 WL 913245, at *4 (S.D.N.Y. Mar. 19, 2012). See Aditi Bagchi, *Other People's Contracts*, 32 YALE J. REG. 211 (2015) for a discussion on reasonable intentions and expectations.

248. *See, e.g.*, Douglas v. U.S. Dist. Ct. for the Cent. Dist. of Cal., 495 F.3d 1062, 1066 (9th Cir. 2007) ("Parties to a contract have no obligation to check the terms on a periodic basis to learn whether they have been changed by the other side. . . . [A]ssent can only be inferred after [the consumer] received proper notice of the proposed changes."); Rodman v. Safeway Inc., No. 11-cv-03003-JST, 2015 WL 604985, at *2 (N.D. Cal. Feb. 12, 2015), *aff'd*, 694 F. App'x 612 (9th Cir. 2017). The court found that, despite Safeway ToS explicitly stating that the company "reserves the right to, from time to time, with or without notice to you in

Similarly, contractual restrictions on access to data can and should be interpreted narrowly. First, since platforms have no proprietary claims to the data itself, researchers may reasonably intend and expect to be able to access and collect platform data once they accept the platform's ToS.[249] Second, accessing platform data is the primary, and often the sole, reason for researchers to use the services offered by digital platforms. Thus, a broad interpretation of platforms' ToS that restricts access to data, even when sought for scientific purposes, would frustrate the researchers' objective for entering into agreements. Third and finally, the power imbalance between platforms and researchers, together with the lack of alternative datasets of equal (or similar) scientific value, imply that, when interpreting agreements between platforms and researchers, courts should apply the interpretation against the drafter rule, which warrants a narrow interpretation of terms restricting access to data to the exclusion of access to data for scientific purposes.

In fact, there are hardly any convincing countervailing reasons to prefer a broad interpretation of platforms' restrictions on data access. As noted, the two primary reasons for (reasonable) platforms to restrict access to data is to advance their business interests and protect users' privacy.[250] However, when access to data is made for scientific purposes, it is unlikely to conflict with platforms' business interests, such as the monetization of data or the use of data to improve and develop competing products and services. Moreover, researchers are already subject to stringent ethical rules, including privacy protection requirements, when they conduct their research, diminishing the second claimed reason for data lockout. Specifically, researchers are required to gain an ethics committee's (Helsinki) approval before commencing a study involving human subjects.[251] Moreover, when applying for funding, researchers conducting data-rich studies are often required to provide a "data management plan" to ensure participants' privacy.[252] Ethically conducted research, therefore, is unlikely to conflict with the users' privacy or with any of their

---

[Safeway's] sole discretion, amend the Terms and Conditions," prior notice is required for changes to be binding.

249. By "reasonable" expectations, we mean the reasonable*,* but not necessarily actual, expectation of researchers. Otherwise, a race to the bottom in firms' drafting practices might arise. For similar concerns pertaining to consumers' expectations in the context of product liability, see Barker v. Lull Eng'g Co., 573 P.2d 443, 454 (Cal. 1978); Douglas A. Kysar, *The Expectations of Consumers*, 103 COLUM. L. REV. 1700, 1749 (2003).

250. See *supra* Section II.A.

251. *See, e.g.*, *WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants*, WORLD MED. ASS'N (Sept. 6, 2022), https://wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [https://perma.cc/DLD5-22FY].

252. *Final NIH Policy for Data Management and Sharing*, NIH (Jan. 25, 2023), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html [https://perma.cc/9L2T-CMMS]; *Data Management Plans*, LONGWOOD RSCH. DATA MGMT. (Jan. 25, 2023), https://datamanagement.hms.harvard.edu/plan-design/data-management-plans [https://perma.cc/6LGU-4D8W] ("A data management plan, or DMP, is a formal document that outlines how data will be handled during and after a research project. Many funding agencies, especially government sources, require a DMP as part of their application processes." (emphasis omitted)).

other interests, implying that there are no reasonable reasons for the parties to intend a broad interpretation of the terms restricting access to platform data.

Still, contract law sometimes permits parties to advance unreasonable objectives.[253] Even so, the parties' intentions, interests, and expectations are not the sole basis for contract interpretation. Courts may also consider the effect of contracts on third parties (i.e., externalities).[254] Specifically, courts will not only refuse to enforce illegal contracts or contracts that are contrary to public policy,[255] but they will also take the public's interests into account when interpreting contracts.[256] Interestingly, the Restatement (Second) of Contracts' sole illustration of interpretation in favor of the public is instructive for our purpose. In that example, a term in an employment termination agreement states that the employee agrees to the assignment of all rights in "a pending patent application and all improvements on the invention covered" to the employer.[257] Though seemingly unrestricted in scope, the Restatement finds that "[t]he public interest in encouraging invention supports an interpretation of the agreement excluding future improvements unless future improvements were specifically included."[258]

Following a similar rationale, a broad interpretation of terms restricting access to data will result in data lockout and elicit the negative externalities of hindering scientific research. Conversely, a narrow interpretation, one allowing access to data for scientific purposes, advances the public interest in increasing knowledge and encouraging invention. It also mitigates the effects of "information monopolies" and provides the means to hold them accountable. Courts have already found these interests to prevail over concerns to users' privacy, even when the entity seeking to access the data is a commercial firm intending to use the data to develop and offer services which compete with those offered by the platform itself.[259] As these countervailing concerns do not pertain to our discussion, we find the case for a narrow interpretation of terms limiting access to data to be even stronger when applied to data access for scientific purposes.[260]

---

253. RESTATEMENT (SECOND) OF CONTRACTS § 203 cmt. c (AM. L. INST. 1981).

254. *See generally* Bagchi, *supra* note 247.

255. *See* Jonathan A. Marcantel, *The Crumbled Difference Between Legal and Illegal Arbitration Awards:* Hall Street Associates *and the Waning Public Policy Exception*, 14 FORDHAM J. CORP. & FIN. L. 597, 597–98 (2009). Marcantel explores how courts will find terms unenforceable for violating public policy through the "public policy exception," which is "a judicial construct prohibiting courts from enforcing illegal contracts or contracts that, while not illegal per se, are against public interests." *Id.*

256. RESTATEMENT (SECOND) OF CONTRACTS § 207 (AM. L. INST. 1981); *see also* Proprietors of Charles River Bridge v. Proprietors of Warren Bridge, 36 U.S. 420 (1837); Larson v. South Dakota, 278 U.S. 429 (1929); Atlanta Ctr. Ltd. v. Hilton Hotels Corp., 848 F.2d 146 (11th Cir. 1988); State Farm Mut. Auto. Ins. Co. v. Est. of Carey, 68 A.3d 1242 (Me. 2012).

257. RESTATEMENT (SECOND) OF CONTRACTS § 207 (AM. L. INST. 1981).

258. *Id.*

259. *See* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1190, 1202 (9th Cir. 2022).

260. Public policy may offer researchers with another avenue of redress. In "Hushing Contracts," David Hoffman and Erik Lampmann argue that, though uncommon in practice, courts should use the contractual public policy doctrine to impede contractual restrictions on access to information pertaining to sexual wrongdoing. Like barriers to platform data, NDAs

2. Unconscionability and Contractual Barriers to Platform Data

The unconscionability doctrine may offer researchers an additional defense against platforms' breach-of-contract claims.[261] The unconscionability doctrine applies to terms that are procedurally and substantively unconscionable,[262] with the two prongs balanced on a sliding scale,[263] and allows courts to "limit the application of any unconscionable clause as to avoid any unconscionable result."[264] Procedural unconscionability pertains to the absence of "real negotiation" and "meaningful choice," which often arise from an imbalance in knowledge, sophistication (understanding), and bargaining power between the parties.[265]

Generally, in consumer contracts, particularly those of digital platforms, procedural unconscionability is manifest. Here, consumers are typically presented

---

attached to settlement agreements undermine the public interest in holding offenders accountable and preventing them from wrongdoing third parties. This, Hoffman and Lampmann argue, justifies such NDAs' unenforceability on public policy grounds. *See* David A. Hoffman & Erik Lampmann, *Hushing Contracts*, 97 WASH. U. L. REV. 165, 167–71 (2019). Were courts to follow Hoffman and Lampmann's approach, the public's interest in platform accountability and the advancement of knowledge implies a similar conclusion as to the (un)enforceability of ToS clauses restricting data access for scientific purposes.

261. That is despite courts at the time being reluctant to apply the doctrine to IP-related disputes. *See, e.g.*, Elkin-Koren, *A Public-Regarding Approach*, *supra* note 147, at 200 (arguing that unconscionability and similar contractual doctrine "are likely to offer only limited help in policing restrictive terms").

262. *See* RESTATEMENT (SECOND) OF CONTRACTS § 208 (AM. L. INST. 1981); Arthur Allen Leff, *Unconscionability and the Code—The Emperor's New Clause*, 115 U. PA. L. REV. 485, 487–88 (1967); Anne Fleming, *The Rise and Fall of Unconscionability as the "Law of the Poor"*, 102 GEO. L.J. 1383, 1423 (2014) ("Leff's distinction—between procedural and substantive unconscionability—has dominated thinking about the issue ever since.").

263. *See, e.g.*, Nagrampa v. MailCoups, Inc., 469 F.3d 1257, 1280 (9th Cir. 2006) ("[T]he more substantively oppressive the contract term, the less evidence of procedural unconscionability is required to come to the conclusion that the term is unenforceable . . . ." (quoting Armendariz v. Found. Health Psychcare Servs., Inc., 6 P.3d 669, 767–68 (Cal. 2000))); E. ALLEN FARNSWORTH, CONTRACTS 302 (3d ed. 2004); RESTATEMENT OF THE LAW, CONSUMER CONTRACTS § 6 cmt. 2 (AM. L. INST. 2022) ("Sliding scale. When both the substantive and the procedural prongs are necessary for a finding of unconscionability, they need not be present in the same degree.").

264. U.C.C. § 2-302(1) (AM. L. INST. & UNIF. L. COMM'N 2003).

265. *See* Williams v. Walker-Thomas Furniture Co., 350 F.2d 445, 449 (D.C. Cir. 1965) ("Unconscionability has generally been recognized to include an absence of meaningful choice on the part of one of the parties together with contract terms which are unreasonably favorable to the other party."); Abramson v. Juniper Networks, Inc., 9 Cal. Rptr. 3d 422, 436 (Ct. App. 2004) ("The oppression component arises from an inequality of bargaining power of the parties to the contract and an absence of real negotiation or a meaningful choice on the part of the weaker party." (internal quotations and citation omitted)); RESTATEMENT OF THE LAW, CONSUMER CONTRACTS § 6(b)(2) (AM. L. INST. 2022) ("[P]rocedural unconscionability, namely a contract or term that results in unfair surprise or results from the absence of meaningful choice on the part of the consumer."). For further discussion on unconscionability, see David Gilo & Ariel Porat, *Viewing Unconscionability Through a Market Lens*, 52 WM. & MARY L. REV. 133, 179–82 (2010).

with a take it or leave it offer by a sophisticated, powerful, and highly informed firm via an agreement spanning several pages and written in technical (legal) language. Furthermore, since digital platforms control the only access points to data of unique scientific value, researchers have no genuine alternative to accessing platform data, further increasing the imbalance between the two parties.[266] Therefore, courts should be able to find procedural unconscionability in agreements between researchers and platforms.

Showing substantive unconscionability, however, may prove more difficult. Substantive unconscionability pertains to the fairness of the contractual terms and is found when terms reflect "one-sidedness" and "gross disparity in the values exchanged."[267] Courts seem especially reluctant to find substantive unconscionability where consumers are offered services for no monetary payment.[268] Commentators, however, have repeatedly called upon courts to rethink their approach and recognize that users "pay" platforms with their time, attention, and effort in the consumption and creation of content. These considerations, to be clear, are not mere "costs" or "detriments" that consumers take upon themselves, but instead they are bargained-for considerations[269]—necessary components of platforms' business models needed to make platforms attractive to advertisers and users alike and fundamental for the creation of the vast amount of data platforms collect, monetize, and use.[270]

Were courts to adhere to these calls, researchers would need to show that terms restricting access to platform data by anyone and for any purpose are one-sided and unfair. Though still not an easy task, several arguments support such a claim. First, as mentioned, reasonable platforms have no reason to impose such absolute restrictions on access to data, as none of their or their users' interests are compromised when researchers access data for scientific purposes. Second, following the discussion of preemption in Section III.A, platforms' use of boilerplate agreements to extend their rights beyond those provided by IP law may be deemed substantively unconscionable, especially where there are no meaningful alternatives

---

266. *See* STEPHEN A. SMITH, CONTRACT THEORY 348–50 (2004).

267. RESTATEMENT (SECOND) OF CONTRACTS § 208 cmt. c (AM. L. INST. 1981); *see* Banner Health v. Med. Sav. Ins. Co., 163 P.3d 1096, 1109 (Ariz. 2007); *Abramson*, 9 Cal. Rptr. 3d at 436 ("Substantively unconscionable terms may take various forms, but may generally be described as unfairly one-sided." (quotation and citation omitted)); Kinney v. United Healthcare Servs., Inc., 83 Cal. Rptr. 2d 348, 353 (Ct. App. 1999) ("'Substantive unconscionability' focuses on the terms of the agreement and whether those terms are 'so one-sided as to shock the conscience.'" (emphasis omitted) (quoting Am. Software, Inc. v. Ali, 54 Cal. Rptr. 2d 477, 482 (Ct. App. 1996))).

268. *See, e.g.*, Song fi, Inc. v. Google Inc., 72 F. Supp. 3d 53, 64 (D.D.C. 2014) ("Having taken advantage of YouTube's free services, Plaintiffs cannot complain that the terms allowing them to do so are unenforceable."). For further discussion see, for example, Bar On, *supra* note 145, at 643–45.

269. *See, e.g.*, *Terms of Service 7*, TIKTOK, https://www.tiktok.com/legal/page/us/terms-of-service/en [https://perma.cc/2QS4-DHH7] (Nov. 2023) ("[B]y submitting User Content via the Services, you hereby grant us an unconditional irrevocable, non-exclusive, royalty-free, fully transferable, perpetual worldwide licence to use, modify, adapt, reproduce, make derivative works of . . . your User Content . . . .").

270. *See* Bar On, *supra* note 145, at 644–45.

for accessing data of similar scientific value, for example, because all other platforms include access-restricting terms in their ToS.[271] Third and finally, the one-sidedness of these terms may also be shown by the fact that the same ToS that restrict researchers from using platform data for scientific purposes also allow platforms to use that data for their own research purposes.[272]

3. Nuisance, Enclosure, and Technological Barriers to Platform Data

Researchers using platform data for scientific purposes can rely on both property-based and contractual-based defenses against potential breach-of-contract claims by digital platforms. We end this Section by suggesting that the common law may also allow scientists to require platforms to remove technological barriers that hinder access to the data. Though one is generally under no obligation to advance the interest of another, the common law recognizes several exceptions to that rule. Even if each example may be debated, one may consider the duty of good faith in performance,[273] disclosure rules,[274] and the fair use doctrine itself as instances of such limitation.

Another such exception comes in the context of access to public resources. Public beaches, for example, are held in trust by the State in favor of the public.[275] Access to beaches, however, is often restricted by beachfront property owners.[276] In response, courts in several states created a right of public access, limiting property owners' right to exclude others. In *State ex rel. Thornton v. Hay*,[277] for example, the Oregon court stated that the public had acquired "an easement for recreational purposes to go upon and enjoy" the public part of the beach all "along the Pacific shore."[278] Similarly, the New Jersey Supreme Court obliged beachfront property owners to provide the public with reasonable access to the beach through their property.[279]

Rights of access were also recognized in the context of public lands. In *Camfield*, a private landowner erected a fence that enclosed the adjoining parcel of public land.[280] After a dispute arose between the landowner and the Bureau of Land

---

271. *See* Fiesler et al., *supra* note 59, at 188–89.

272. *See, e.g.*, *Privacy Policy*, META (Dec. 27, 2023), https://www.facebook.com/privacy/policy/version/7122790421067234/ [https://perma.cc/XJ59-43QX] ("We use information we have, information from researchers and datasets from publicly available sources, professional groups and non-profit groups to conduct and support research.").

273. *See, e.g.*, Market Street Assocs. Ltd. v. Frey, 941 F.2d 588, 597 (7th Cir. 1991) ("[D]eliberately to take advantage of your contracting partner's mistake during the performance stage . . . is a breach of good faith . . . .").

274. Laidlaw v. Organ, 15 U.S. 178, 193–94 (1817).

275. *See* Ill. Cent. R.R. v. Illinois, 146 U.S. 387 (1892).

276. *See* James M. Kehoe, *The Next Wave in Public Beach Access: Removal of States as Trustees of Public Trust Properties*, 63 FORDHAM L. REV. 1913, 1913–17 (1995).

277. 462 P.2d 671 (Or. 1969).

278. *Id.* at 673–76; *see also In re* Ashford, 440 P.2d 76, 77–78 (Haw. 1968).

279. *See* Raleigh Ave. Beach Ass'n v. Atlantis Beach Club, Inc., 879 A.2d 112, 113–25 (N.J. 2005).

280. Camfield v. United States, 167 U.S. 519 (1897).

Management staff managing it, the Supreme Court ordered the removal of the fence, finding that because it was "intended to enclose the lands of the [g]overnment," it constituted a nuisance, despite being entirely on private lands.[281]

Camfield, to be sure, discussed the enclosure of public lands which, unlike data, are owned by the Federal Government. But the rationale for providing right of access to beaches and public land should equally apply to access to *non-proprietary* data, such as UPD. Early signs of courts adopting such approach can be found in *hiQ*. In that case, *hiQ* used publicly available UPD hosted on LinkedIn's servers for commercial purposes. In response, "LinkedIn sent hiQ a cease-and-desist letter, asserting that hiQ was in violation of LinkedIn's User Agreement" as well as of state and federal law, and stated that "LinkedIn had 'implemented technical measures to prevent hiQ from accessing LinkedIn's site . . . and block scraping activity.'"[282] Affirming the district court's decision, the Court of Appeals for the Ninth Circuit granted hiQ's request for preliminary injunctive relief, ordering LinkedIn to "[w]ithdraw its cease-and-desist letter, to remove any existing technical barriers to hiQ's access to public profiles, and to refrain from putting in place any legal or technical measures with the effect of blocking hiQ's access to public profiles."[283] In effect, the court treated both the legal and the technological measures LinkedIn used to deny access to publicly available UPD as a psychical fence enclosing public lands.

Moreover, when considering LinkedIn's claim that allowing hiQ and others like it unrestrained access to platform data would put users' privacy at risk, the court found that, though there is a clear public interest in protecting users' privacy, the public will be better served if access to data is granted. "[G]iving companies like LinkedIn free rein to decide, on any basis, who can collect and use data—data that the companies do not own, that they otherwise make publicly available to viewers, and that the companies themselves collect and use," the court reasoned, "risks the possible creation of information monopolies that would disserve the public interest."[284]

The *hiQ* decision was given in the context of access to data for competing, commercial purposes. The case of access to data for scientific, noncommercial, and noncompeting purposes is arguably more compelling. Here, the public's interest is not only in limiting the power of "information monopolies," as the court put it, but also in advancing human knowledge. Moreover, when access is made by researchers, who must meet stringent ethical requirements,[285] privacy concerns are largely mitigated. That the public interest is further advanced when access to platform data is sought for scientific purposes, as opposed to commercial ones, should inform the scope of the obligation to remove legal and technological barriers. For example, because it better serves the public interest, such obligation may extend beyond publicly available PPD, and into non-publicly available UPD and CPD.

---

281. *Id.* at 528; *see also* Leo Sheep Co. v. United States, 440 U.S. 668, 685–86 (1979) (finding that, while *Camfield* requires the removal of barriers to access, it does not provide an affirmative right of public access).

282. hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1187 (9th Cir. 2022).

283. *Id.* at 1188.

284. *Id.* at 1202.

285. *See supra* note 251 and accompanying text.

C. From Common Law to Regulation: An Affirmative Right to Access Platform Data

Private law doctrines, as discussed above, may serve as a shield to researchers from potential legal liability when accessing platform data for scientific purposes. However, while such a shield may be crucial for defending against claims alleging violation of the platform terms and conditions, it may not provide sufficient assurances to scientists relying on platform data for their studies. This is due to several reasons: First, scientists may face a high level of uncertainty, as platforms often rely on vaguely and broadly defined provisions in their ToS to restrict research practices.[286] The uncertainty surrounding court interpretation of ToS in specific instances may hence result in a chilling effect on researchers. Second, even if researchers were assured of the legal outcome, the prospect of prolonged litigation against a well-funded repeat actor with much at stake is a significant deterrent.[287] Scientists who choose to utilize legally contested data, banking on the belief that courts would rule in their favor, are not simply confronting legal liability. The successful completion of the research itself might be jeopardized if platforms decide to take action against the publication of research results. Such actions may further subject researchers to publication delays, as leading publishers may be hesitant to publish articles reliant on such data. Given that scientific research is often time sensitive,[288] the prospect of such delays, coupled with the threat of legal liability, may deter scientists from utilizing platform data for scientific purposes.

Finally, a legal shield based on private law doctrines might be most effective with respect to publicly accessible data. Yet, as noted, not all platform data is publicly accessible, and researchers might require cooperation from platforms to access some internal data.[289] Therefore, it is crucial not only to protect researchers from potential legal liability, but also to clearly define their right to access platform data.

This Section explores whether existing or proposed regulatory initiatives could complement private law doctrines to ensure access to platform data for scientific purposes. We first discuss some regulatory initiatives mandating data access and then analyze whether they could substitute for private law doctrines.

1. Mandating Access to Platform Data by Regulation

Regulatory initiatives in Europe and in the United States aim to address barriers to accessing platform data, with the goal of ensuring accountability of online

---

286. *See supra* notes 241–242 and accompanying text.

287. *See* Kayser-Bril, *supra* note 104.

288. Scientific publications are time sensitive due to the accelerated pace of scientific development. Scientists strive to be the first to achieve a scientific milestone, to acquire the reputational benefits, compete for funding and placement in high-ranking scientific outlets, or to simply satisfy promotion requirements.

289. *See supra* Section I.C.

platforms and promoting greater oversight.[290] As discussed above, however, the need to ensure access for independent scientific research goes further than that.[291]

In the case of access for research, the European Union has taken a leadership role by moving to mandate some access to platform data for scientific purposes, in response to a mounting pressure to enable independent studies and oversight.[292] The Digital Services Act (DSA), which became effective in November 2022, mandates digital platforms to facilitate access to data for public interest research in compliance with a newly established governance framework. It aims to streamline access to data, while also addressing the legitimate interests of platforms and their users.

The DSA introduces a novel regulatory body, the Digital Services Coordinators, tasked with the management of data access authorizations. This new agency shifts the decision-making power regarding access to data from the sole discretion of profit-driven platforms to an administrative agency entrusted with upholding the public interest.[293]

The DSA further introduces a structured procedure for obtaining access to data for research purposes, including a filing procedure and eligibility criteria for researchers and their proposed projects.[294]

Most notably, the DSA obliges very large online platforms and search engines (VLOPs and VLOSEs) to provide data to "vetted researchers"[295] for the sole purpose of conducting research that contributes to "the detection, identification and understanding of systemic risks in the Union," as set out pursuant to Article 34(1), and "to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 35."[296]

Meanwhile, in the United States, several bills have been introduced, proposing to bind some digital platforms to make data available for research purposes. Most notably, the Platform Accountability and Transparency Act[297] (PATA), compels

---

290. Brandie Nonnecke & Camille Carlton, *EU and US Legislation Seek to Open Up Digital Platform Data*, 375 SCIENCE 610, 611–12 (2022).

291. *See supra* notes 32–44 and accompanying text.

292. DSA, *supra* note 31, at 27.

> This Regulation therefore provides a framework for compelling access to data from very large online platforms and very large online search engines to vetted researchers affiliated to a research organization within the meaning of Article 2 of Directive (EU) 2019/790, which may include, for the purpose of this Regulation, civil society organisations that are conducting scientific research with the primary goal of supporting their public interest mission.

Id.

293. *See id.* at 79–82 arts. 49–51.

294. *Id.* at 70–72 art. 40.

295. *Id.* at 70–71 art. 40(4), (8).

296. *Id.* at 70 art. 40(4).

297. Platform Accountability and Transparency Act (PATA), S. 1876, 118th Cong. (2023). The bill was first introduced in 2022 and was reintroduced in June 2023 with minor changes. *See* John Perrino, *Platform Accountability and Transparency Act Reintroduced in Senate*, TECH POL'Y PRESS (June 8, 2023), https://www.techpolicy.press/platform-accountability-and-transparency-act-reintroduced-in-senate/ [https://perma.cc/MTZ7-FCN4].

large platforms[298] to make data available to qualified researchers.[299] The bill establishes a process where researchers may require access to certain data from digital platforms. The process is facilitated by the National Science Foundation (NSF) to ensure scientific merits, and by the Federal Trade Commission (FTC) to address privacy and cybersecurity concerns.[300] Under the bill, these bodies would be required to establish a research program for soliciting, reviewing, and approving research proposals and for providing guidelines on conducting and publishing qualified research projects.[301]

Importantly, the bill introduces a safe harbor for researchers or journalists who collect publicly available data from digital platforms, using "covered method of digital investigation" to "inform the general public about matters of public concern," provided they comply with privacy, security, and public interest work requirements.[302]

Other bills similarly aim to facilitate access to platform data for research purposes.[303] For instance, The Kids Online Safety Act provides researchers with access to social media platforms for studying specific harms.[304]

At the same time, however, despite increasing concerns about researchers' lack of access to platforms' data and the growing number of new bills, as of today, none of these bills have been enacted into law.

2. Complementing Regulatory Gaps

Does the regulatory framework ensure access to platform data for scientific purposes? If adopted by the U.S. legislature, would it render the private law strategies for securing access to data redundant? The following discussion demonstrates how these strategies could be complementary.

---

298. S. 1876, § 2(5)(B) ("[H]as at least 50,000,000 unique monthly users in the United States for a majority of the months in the most recent 12-month period.").

299. The bill further includes public transparency requirements to allow public access to advertising libraries and disclosures on viral content. S. 1876.

300. A unit to be established within the Federal Trade Commission, namely, the Platform Accountability and Transparency Office (PATO). S. 1876, § 3 ("Not later than 1 year after the date of enactment of this Act, the NSF shall establish, in consultation with the Commission, a research program to review research applications for approval as qualified research projects.").

301. *Id.* § 3.

302. *Id.* § 8(a) ("No civil claim will lie, nor will any criminal liability accrue, against any person for collecting covered information as part of a news-gathering or research project on a platform, so long as . . . .").

303. For instance, if enacted, the Social Media Disclosure and Transparency (DATA) Act would require certain platforms to maintain advertisement libraries and make them available to academic researchers and the Federal Trade Commission (FTC). The bill further proposes establishing a working group to address social media research access and make policy recommendations regarding the type of data digital platforms should make available to academic researchers. *See* Social Media DATA Act, H.R. 3451, 117th Cong. (2021).

304. Kids Online Safety Act, S. 1409, 118th Cong. (2023).

a. Reducing Uncertainty—but Lowering Flexibility

As discussed above, a major barrier for scientists is not simply the potential legal liability involved in accessing publicly available data on digital platforms, but also the uncertainty regarding the scope of liability and the ability to sustain the research project.[305] Given scientists' dependency on risk-averse academic institutions, funders, and publishers, uncertainty regarding permissible uses of data exacerbated by risk aversion would result in a chilling effect.

Regulation may offer certainty regarding issues such as data access application procedures and the grounds for rejecting a specific data access application. At the same time, however, many questions still remain open. Even a detailed regulation such as the DSA has some major gaps. For instance, does Article 40, or its equivalent future procedure under the PATA bill, include only a duty to disclose data or also the ability to conduct experiments? Are exploratory studies, which lack definitive research questions in the outset, allowed? Does the duty to disclose data also imply the provision of metadata regarding any dataset acquired under such procedure? Which methodologies of extracting data are allowed? For instance, does the law permit scraping from publicly available web pages? Or how will the duties imposed by regulation for data security and privacy protection be implemented?[306]

Indeed, the EU Commission has recently launched a call for evidence on the DSA related to data access for research purposes intended to inform the implementation of Article 40.[307] Respondents to this call have stressed the need to provide standard procedures and criteria for eligibility to vetted researchers to establish an independent advisory body with professional expertise and address liability for potential data breaches. Based on the contributions received, the Commission is scheduled to prepare a delegated act on Article 40 to be adopted in 2024.[308]

Gaps in any regulation are unavoidable, however. It is impossible to predict and address ex ante all the possible needs, types of data, methodologies, and circumstances of data reused for research purposes. New emerging needs would require a more nuanced adjustment process, which might be more consistent with common law practices of case-by-case interpretation.

b. Scientific Access Only as Instrumental to Accountability

A major shortcoming of the current legislative efforts is that they are clearly motivated by an objective that is not directly related to promoting scientific goals, but rather at promoting platform accountability and transparency.[309] For instance,

---

305. *See supra* notes 286–289 and accompanying text.

306. Daphne Keller, Commentary, *Delegated Regulation on Data Access Provided for in the Digital Services Act*, 2023 STAN. CYBER POL'Y CTR. 1, 2–6.

307. *Digital Services Act: Summary Report on the Call for Evidence on the Delegated Regulation on Data Access*, EUR. COMM'N (Nov. 24, 2023), https://digital-strategy.ec.europa.eu/en/library/digital-services-act-summary-report-call-evidence-delegated-regulation-data-access [https://perma.cc/2REL-XNJH].

308. *Id.*

309. For instance, a recent Congressional report states that "[m]any academic and third-party researchers lack access to the internal data, models, and other information that could be

eligible research under the DSA[310] is defined as research which "contributes to the detection, identification and understanding of systemic risks in the Union," and to the assessment of the adequacy, effectiveness, and impact of risk mitigation measures.[311] By focusing on enhancing platform accountability and oversight, the DSA may fail to facilitate general scientific research. The proposed U.S. PETA Bill defines the purpose of eligible research more broadly—that is "to inform the general public about matters of public concern"[312]—yet the bill as a whole is intended to facilitate access to data to promote accountability and transparency.

This narrow view may shape research agendas of academic researchers by facilitating a narrow subset of studies and at the same time failing to accommodate the barriers, which apply to others. As we have demonstrated elsewhere,[313] such an instrumental view of scientific research—only as a means for promoting platform accountability—overlooks the social benefits arising from independent scientific research and fails to ensure access to data for general-purpose research, which is essential for the promotion of knowledge and for strengthening the independence of scientific enterprises.[314] For example, the narrow view may exclude scientific uses not directly informative to the public, such as the training of NLP models, thereby allowing digital platforms to maintain their status as information monopolies in these contexts.

The shortcomings of the narrowly defined, legitimate objectives of prospective studies are also coupled with the general concern that a regulatory framework, while reducing uncertainty, would not be sufficiently flexible to accommodate the fast-moving practices of conducting scientific research.

---

needed to conduct comprehensive studies of online platforms. These studies could provide insight into online platforms and their effects on users." CLARE Y. CHO & LING ZHU, CONG. RSCH. SERV., R47662, DEFINING AND REGULATING ONLINE PLATFORMS 20 (2023).

310. DSA, *supra* note 31, at 27, 72 art. 40(12). DSA, art. 40(12) and Recital 98.

311. *Id.* at 70 art. 40(4). Arguably, systemic risks and risk mitigation measures can be interpreted broadly to cover different types of studies affecting any fundamental right or freedoms. *See* Paddy Leerssen, *Counting the Days: What to Expect from Risk Assessments and Audits Under the DSA – and When?*, DSA OBSERVATORY (Jan. 30, 2023), https://dsa-observatory.eu/2023/01/30/counting-the-days-what-to-expect-from-risk-assessments-and-audits-under-the-dsa-and-when [https://perma.cc/NUM3-33TL]. Yet by focusing research that investigates the impact of *digital platforms* on fundamental rights, the DSA leaves out research intended to identify general risks to fundamental rights.

312. Platform Accountability and Transparency Act (PATA), S. 1876, 118th Cong. § 8 (2023) (creating a safe harbor for platform data research, as long as "the purpose of the project is to inform the general public about matters of public concern").

313. Aline Iramina, Maayan Perel & Niva Elkin-Koren, *Paving the Way for the Right to Research Platform Data*, SSRN (June 19, 2023), https://ssrn.com/abstract=4484052 [https://perma.cc/CGV6-DAH6].

314. *Id.* at 14–16; *see also* Keller, *supra* note 306; Paddy Leerssen, *Call for Evidence on the Delegated Regulation on Data Access Provided for in the Digital Services Act - Summary & Analysis*, UNIV. OF AMSTERDAM (DIGIT. ACAD. REPOSITORY) (2023), https://pure.uva.nl/ws/files/155412569/DSA_Data_Access_Call_for_evidence_summary_2alttext_2deaRi2SZpXsUUwmL5Y17seUl9g_100331.pdf [https://perma.cc/85RL-D5YW].

c. A Right to Research

While regulation such as the DSA may still have some imperfections, it marks a significant stride toward establishing a legal right for researchers to request access to platform data and puts in place an institutional framework to facilitate the exercise of this right.[315] Indeed, private law may sometimes create new duties, yet it remains highly contested to what extent private law should impose an affirmative duty on one person to advance the project of another.[316] This suggests that complementary measures are necessary so that private law can accommodate researchers' access to platform data, equipping them not only with a shield, but also with a sword.

By introducing this legal obligation, the DSA effectively establishes a (limited) right to conduct academic research on the systemic risks involving digital platforms in the EU. This right encompasses the ability to request data collection, use APIs, or engage in other means of automatic extraction. This might be crucial for conducting research in the digital environment, especially when access is required to internal data that is not publicly available.

By establishing duties of platforms to provide access to data and the correlative right of researchers to access platform data, the regulation also provides a normative signal. Recognizing this right could inform the courts' interpretation of platforms' ToS. For instance, in circumstances where platforms seek to contractually restrict scraping of certain publicly available data.

Thus, establishing a right to research may not simply create a sword for requiring access to otherwise unavailable platform data but may also strengthen the shield of scientists against contractual claims.

CONCLUSION

Platform data is fueling scientific progress in the digital era. In the search for new insights about humans, markets, and various scientific challenges, data-driven research methodologies take the front seat. It is therefore becoming critical to both facilitate and ensure researchers' access to platform data. Policymakers, both in and outside the United States, seem to be attuned to this important need, seeking to mandate digital platforms to cooperate with researchers and share with them data that is needed to oversee platforms and hold them accountable. Reforms of this sort are important to support platform data research because they grant researchers an affirmative right of access. Yet, as this Article demonstrates, more is needed to adequately facilitate the full range of potential platform data research. As much as researchers need a clear and certain mechanism for obtaining access to platform data, they also need comprehensive legal defenses against breach-of-contract claims that may be asserted by well-funded platforms. As demonstrated in this Article,

---

315. Iramina et al., *supra* note 313.

316. Though a highly contested issue, right-based theories, such as corrective justice, generally reject the use of private law to advance public interests; consequentialist theories of law, such as economic analysis, may take the public interest in account in their legal analysis of legal rules (e.g., when considering their impact on overall social welfare). *See* Peter Benson, *The Idea of a Public Basis of Justification for Contract*, 33 OSGOODE HALL L.J. 273, 328–29 (1995).

contractual restrictions effectively govern data access on platforms in a robust and unilateral manner. Therefore, to protect researchers from potential liability, this Article illustrates how private law defenses, rooted in copyright law, contracts law, and the doctrine of nuisance, can enable researchers to effectively contest breach-of-contract claims by powerful platforms. Considering the rapid pace of technological change, which directly influences research methodologies, the collection of data by platforms, and the dynamic nature of platforms' ToS, a flexible legal shield may be researchers' most essential requirement in the long run.