THE DELETION REMEDY[*]

DANIEL WILF-TOWNSEND[**]

A new remedy has emerged in the world of technology governance. When someone has wrongfully obtained or used data, this remedy requires them to delete not only that data but also to delete tools such as machine learning models that they have created using the data. Model deletion, also called algorithmic disgorgement or algorithmic destruction, has been increasingly sought in both private litigation and public enforcement actions. As its proponents note, model deletion can improve the regulation of privacy, intellectual property, and artificial intelligence by providing more effective deterrence and better management of ongoing harms.

But, this Article argues, model deletion has a serious flaw. In its current form, it has the possibility of being a grossly disproportionate penalty. Model deletion requires the destruction of models whose training included illicit data in any degree, with no consideration of how much (or even whether) that data contributed to any wrongful gains or ongoing harms. Model deletion could thereby cause unjust losses in litigation and chill useful technologies.

This Article works toward a well-balanced doctrine of model deletion by building on the remedy's equitable origins. It identifies how traditional considerations in equity—such as a defendant's knowledge and culpability, the balance of the hardships, and the availability of more tailored alternatives—can be applied in model deletion cases to mitigate problems of disproportionality. By accounting for proportionality, courts and agencies can develop a doctrine of model deletion that takes advantage of its benefits while limiting its potential excesses.

[**] Associate Professor of Law, Georgetown University Law Center. I would like to thank David Bau, Anupam Chander, Paul Ohm, Tanina Rostain, Alicia Solow-Niederman, and Neel Sukhatme for helpful conversations and suggestions. Thanks also to Erika Sloan for excellent research assistance, and to the editors of the *North Carolina Law Review* for their careful attention and thoughtful improvements.

INTRODUCTION

What should the law do when a company obtains, shares, or uses data unlawfully? A standard answer is to fine the company—make it pay some money, as is typical with companies that violate the law.[1] And maybe the company should have to destroy its copies of the data, too—it might be strange to allow the company to keep the data, particularly if the violation involved possessing it in the first place.[2] And finally, to decrease the odds of these problems happening again in the future, it may make sense to enter some kind of enforceable agreement stipulating ways in which the company will change its practices going forward.[3] For decades, these three types of relief have been the

[1] *See, e.g.*, Press Release, FTC, FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook (July 24, 2019), https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook [https://perma.cc/HKA6-Z26C].

[2] *See, e.g.*, Press Release, FTC, FTC and DOJ Charge Amazon with Violating Children's Privacy Law by Keeping Kids' Alexa Voice Recordings Forever and Undermining Parents' Deletion Requests (May 31, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-doj-charge-amazon-violating-childrens-privacy-law-keeping-kids-alexa-voice-recordings-forever [https://perma.cc/2VNX-R2ZZ].

[3] *See, e.g.*, Stipulated Order for Civil Penalty, Monetary Judgment, and Injunctive Relief, United States v. Facebook, Inc., No. 19-cv-2184 (D.D.C. July 24, 2019).

defining remedies in cases where data has been obtained, possessed, or used in ways that violate the law.[4]

But in recent years, both private and public litigants have hit on a new remedy: destroying work product that is made from the data at issue. Data is often valuable not just in its own right but because it is useful for building something: a model that can make predictions or classifications, or generate images, audio, or text.[5] These models, in turn, can survive and be used even after the data that was used to train them has been destroyed.[6] Paying attention only to the data, and not to the models derived from the data, may therefore be inadequate when it comes to deterring misconduct or preventing ongoing harm that arises from these models' use.[7]

Seeing this problem, the Federal Trade Commission ("FTC") pioneered the remedy of model deletion in its investigation of Cambridge Analytica in 2019.[8] Model deletion—requiring the deletion of models trained on unlawfully used or possessed data—has since caught on, appearing regularly in FTC orders in recent years, and moving into the repertoire of other public enforcers as well as private litigants.[9] Along the way, model deletion has drawn considerable attention and praise from public officials and civil society.[10] Model deletion has

[4] *See, e.g.*, FTC, 40 YEARS OF EXPERIENCE WITH THE FAIR CREDIT REPORTING ACT 3–5, 108–10 (2011), https://www.ftc.gov/sites/default/files/documents/reports/40-years-experience-fair-credit-reporting-act-ftc-staff-report-summary-interpretations/110720fcrareport.pdf [https://perma.cc/VGC7-48XA] (collecting enforcement actions under the Fair Credit Reporting Act).

[5] *See* Amanda Parsons & Salomé Viljoen, *Valuing Social Data*, 124 COLUM. L. REV. 993, 1026–35 (2024).

[6] *See infra* Section I.B.

[7] *See* Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. REV. 479, 496–98 (2022) (discussing the limitations of a remedial approach that focuses only on data deletion and not model deletion); Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 NW. U. L. REV. 357, 400–03 (2022) [hereinafter Solow-Niederman, *Inference Economy*] (discussing harms that can arise from machine learning tools' inferential capacities).

[8] Final Order at 4, Cambridge Analytica, LLC, F.T.C. Docket No. 9383 (Nov. 25, 2019) (consent order).

[9] *See infra* Section I.B.

[10] *See* Li, *supra* note 7, at 493–96; Christina Lee, *Beyond Algorithmic Disgorgement: Remedying Algorithmic Harms*, 16 U.C. IRVINE L. REV. (forthcoming 2026) (manuscript at 5–7); Rebecca Kelly Slaughter, Janice Kopec & Mohamad Batal, *Algorithms and Economic Justice: A Taxonomy of Harms and A Path Forward for the Federal Trade Commission*, 23 YALE J.L. & TECH. 1, 5–6 (2021); Jevan Hutson & Ben Winters, *America's Next "Stop Model!": Model Deletion*, 8 GEO. L. TECH. REV. 124, 127 (2024); Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data*, 29 RICH. J.L. & TECH. 1, 39–47 (2023); Emma Elder, *Wrongful Improvers as a Guiding Principle for Application of the FTC's IP Deletion Requirement*, 97 WASH. L. REV.

been called "an innovative and promising remedy,"[11] and a "unique and viable enforcement option."[12] Commenters have said it reflects "a more sophisticated understanding" of how data matters,[13] suggesting it will "motivate companies to get their act together,"[14] and have suggested making it a "privacy right to be included in privacy laws."[15] In short, model deletion appears at first glance to be an ideal remedy: depriving a wrongdoer from benefiting from misconduct, deterring bad actors, and providing a better remedy for ongoing harm than data deletion or fines alone.

But in its current form, model deletion suffers a major flaw: it has the potential to be a grossly disproportionate remedy.[16] As formulated by the FTC, model deletion requires a defendant to "[d]elete . . . any information or work product, including any algorithms or equations, that originated, in whole or in part, from" data that was unlawfully obtained, possessed, or used.[17] This approach amounts to a "no bad bytes" rule, which means that any software model must be destroyed if it was created out of a dataset where any fraction of the data was unlawful in some way. There does not need to be a connection between the unlawful data and the model's purpose or use: the model could be used in ways that are unrelated to the unlawful portion of its training data, and model deletion would still require it to be destroyed. Nor does model deletion require any assessment of a defendant's culpability, such as examining their knowledge or intentionality with respect to the unlawful data at issue. And there is no inherent requirement for an examination of the value of the model being destroyed, or any comparison of that value to the costs or harms of the unlawful

1009, 1013–14 (2022); Solow-Niederman, *Inference Economy*, *supra* note 7, at 377–78; *see also* Pamela Samuelson, *How to Think About Remedies in the Generative AI Copyright Cases*, LAWFARE (Feb. 15, 2024, 1:00 PM), https://www.lawfaremedia.org/article/how-to-think-about-remedies-in-the-generative-ai-copyright-cases [https://perma.cc/UQD9-RC3N] (discussing the possibility of model deletion in the context of copyright litigation without making claims about the remedy more broadly).

[11] Slaughter et al., *supra* note 10, at 5–6.

[12] Hutson & Winters, *supra* note 10, at 127.

[13] Solow-Niederman, *Inference Economy*, *supra* note 7, at 377.

[14] Tonya Riley, *The FTC's Biggest AI Enforcement Tool? Forcing Companies to Delete Their Algorithms*, CYBERSCOOP (July 5, 2023), https://cyberscoop.com/ftc-algorithm-disgorgement-ai-regulation/ [https://perma.cc/R7H3-29PQ] (internal quotation marks and brackets omitted).

[15] Li, *supra* note 7, at 504.

[16] *See infra* Part III.

[17] Final Order at 4, *Cambridge Analytica, LLC*, *supra* note 8. This language is not identical in every FTC order, but so far, all orders for model deletion have used similar language. *See infra* Section I.B.

possession or use of the data at issue.[18] The result is a remedy that is promising, but one that has significant latent flaws.

These flaws have been obscured because model deletion has so far only been deployed in a series of FTC actions where it was plausibly reasonable and proportionate to use the remedy.[19] But as the commercial use of gigantic models trained on vast datasets continues to grow, the likelihood—and potential downsides—of a disproportionate use of model deletion grows correspondingly.[20] And as private litigants and a broader range of public enforcers begin to seek out model deletion as a remedy, it is implausible to rely on the enforcement discretion of those seeking model deletion to stay the inherent power of this remedial tool. To the contrary, the availability of model deletion could easily act as a perverse incentive in private litigation, turning claims that would otherwise be small and easily settled into high-stakes, bet-the-company litigation, as a single loss could result in the destruction of some companies' most valuable assets.[21]

These problems arise in part because there is not yet any meaningful law or doctrine around model deletion. So far, the legal bases invoked to support model deletion have been statutory provisions granting courts and agencies broad equitable authority.[22] No statute or regulation specific to model deletion exists. No court has weighed in on model deletion yet, nor has any public agency developed a robust legal justification or guidelines for the remedy.[23] As a result, there are basically no contours to model deletion—no set of factors or guidelines to consider when evaluating whether the remedy is appropriate. Instead, there is only a handful of enforcement actions introducing a remedial tool that is reasonable in some circumstances but that could result in grossly disproportionate penalties in others.

Many justify the deletion remedy by comparing it to disgorgement. Disgorgement, an equitable remedy, requires wrongdoers to turn over the benefits that they have derived from their bad acts.[24] FTC officials have drawn on this concept for their model deletion orders, referring to model deletion as "algorithmic disgorgement" and framing it as depriving wrongdoers of the

[18] *See infra* Part III.
[19] *See infra* Part III.
[20] *See infra* Part III.
[21] *See infra* Part III.
[22] *See infra* Section III.A.
[23] *See infra* Section I.B.
[24] *See infra* Section III.B.

benefit of their misconduct.[25] But model deletion as the FTC has articulated it so far is not clearly supported by the law or doctrine of disgorgement.[26] Disgorgement is a remedy that is required to be proportioned to the unjust gains at issue and also requires a demonstration that whatever is disgorged is causally linked to the wrongful conduct.[27] But these requirements are not satisfied by model deletion in easy-to-imagine scenarios—such as where a defendant has trained a model on a large dataset, and the unlawful data at issue is neither a

[25] *See, e.g.*, Slaughter et al., *supra* note 10, at 39; Remarks of Samuel Levine, Director, Bureau of Consumer Protection, Federal Trade Commission, at 10–11, Cleveland-Marshall College of Law Cybersecurity and Privacy Protection Conference (May 19, 2022), https://www.ftc.gov/system/files/ftc_gov/pdf/Remarks-Samuel-Levine-Cleveland-Marshall-College-of-Law.pdf [https://perma.cc/YV8U-KDQP] (describing model deletion as resulting from "the simple principle that companies should not be able to profit from illegal data practices"); Statement of Commissioner Rohit Chopra at 1, Everalbum and Paravision, F.T.C. Case No. 1923172 (Jan. 8, 2021) [hereinafter Statement Regarding Everalbum], https://www.ftc.gov/system/files/documents/public_statements/1585858/updated_final_chopra_statement_on_everalbum_for_circulation.pdf [https://perma.cc/4DG7-8VYZ] (describing an FTC model deletion order as "requir[ing] Everalbum to forfeit the fruits of its deception").

The variety of terms used to describe model deletion is symptomatic of a wider problem involving overly broad or unclear terms in the world of AI and the law. *See, e.g.*, David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669–70 (2017). In addition to "algorithmic disgorgement," there are a variety of other names that model deletion has gone by: "algorithmic destruction," "model destruction," "model disgorgement," or "model deletion." *See* Hutson & Winters, *supra* note 10, at 128–29. This Article uses "model deletion." As discussed *infra* Section III.B., the kind of orders at issue here only loosely parallel the concept of disgorgement, and in some contexts, it may be misleading to think of them as supported by disgorgement doctrine.

As for "model" versus "algorithm," they are often used interchangeably. *See*, *e.g.*, Lehr & Ohm, *supra*, at 671. There is a usage pattern in the machine learning context in which "algorithm" is the broader term, encompassing models and more, while a "model" is more specifically the result of an algorithmic machine learning training process. *See, e.g.*, Cathy Petrozzino, *Big Data Analytics*, 16 SCITECH LAW. 14, 15 (Spring 2020) ("[A]n algorithm is comprised of a set of rules that need to be followed in order to solve a problem. A model is built by using an underlying algorithm and is shaped by the training data."); Andrew Amann, *Machine Learning Algorithm or Machine Learning Model?*, TECHOPEDIA (Sept. 13, 2022), https://www.techopedia.com/machine-learning-algorithm-or-machine-learning-model/7/34855 [https://perma.cc/ZX97-LX43 (staff-uploaded archive)]. I follow that pattern in sticking with the phrase "model deletion" over "algorithmic deletion," as the emphasis of this remedy is in destroying the downstream product of training. More recently, the FTC has also begun using the phrase "model deletion." *See* Press Release, FTC, FTC Order Will Ban Avast from Selling Browsing Data for Advertising Purposes, Require It to Pay $16.5 Million Over Charges the Firm Sold Browsing Data After Claiming Its Products Would Block Online Tracking (Feb. 22, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-order-will-ban-avast-selling-browsing-data-advertising-purposes-require-it-pay-165-million-over [https://perma.cc/X2PH-4Z9S].

[26] *See infra* Section III.B.

[27] *See infra* Section III.B.

significant portion of the broader dataset nor a distinctly valuable subset of it.[28] In that circumstance, it could well be the case that the unlawful data is not a cause of much or any of the model's value, nor a cause of any harm attributable to the model.[29] But model deletion's application to models "trained in whole or in part" on unlawful data would result in the deletion of the model. As that kind of case demonstrates, model deletion's current formulation contains none of the safeguards of disgorgement doctrine that are designed to ensure proportionality.

This Article attempts to chart a path forward for using model deletion's distinct advantages as a remedy in a way that is sensitive to its potential problems of disproportionality. It does so by working toward an equitable doctrine of model deletion. Model deletion's grounding in courts' and agencies' equitable powers provides an opportunity to create a flexible, context-specific remedy that is much more functional than its existing uses might make it seem. Using traditional equitable guideposts, courts weighing model deletion as an option can consider a defendant's culpability; compare the potential harms to the plaintiff, the defendant, and third parties; and consider the possibilities of alternative remedies.[30] As a result, the doctrine of model deletion can develop in a productive and fair direction, even without new legislation or regulation.

Getting model deletion right is important. The unlawful use of data is one of the defining governance issues of our time. Developments in technology, law, and social institutions in recent decades have created an economy in which access to data creates profits and market power.[31] The gains from data, which can be massive, can make companies willing to flout regulations—posing a major obstacle to the enforcement of laws limiting the collection, possession, exchange, or use of data and information.[32] It is a common complaint that the United States has done a poor job passing laws to manage the technologies and systems of the present day.[33] It is also true that our institutions often have

[28] *See infra* Section III.B.

[29] *See infra* Section III.B.

[30] *See infra* Part IV.

[31] JULIE E. COHEN, BETWEEN TRUTH AND POWER 25–47 (2019).

[32] *See, e.g.*, Rohit Chopra, *Reining in Repeat Offenders*, 11 REG. REV. DEPTH 9, 14 (describing Facebook's willingness to "openly flout[]" FTC orders).

[33] *See, e.g.*, Jessica Rich, *After 20 Years of Debate, It's Time for Congress to Finally Pass a Baseline Privacy Law*, BROOKINGS (Jan. 14, 2021), https://www.brookings.edu/articles/after-20-years-of-debate-its-time-for-congress-to-finally-pass-a-baseline-privacy-law/ [https://perma.cc/25FY-DY6N].

difficulty enforcing the laws that we already have.[34] A new remedy more tailored to the abuses of data has significant promise for the fair and effective enforcement of our laws.[35]

But the kinds of large data agglomerations that make it profitable for companies to flout the law also contain significant prospects for social welfare. And that means that for model deletion to be a valuable remedy, it must be well-tailored. The predictive insights that can be derived from big data have produced improvements in fields as varied as scientific research, healthcare, engineering, education, and government.[36] These tools, when well-deployed, have the prospect of saving lives and reducing discrimination.[37] Predictive and generative AI, which typically rely on large datasets to train, have led to tens of billions of dollars of investment across many industries, suggesting the possibility of further benefits yet to come.[38] There are plenty of commercial uses of these tools that will have negative consequences—hence the need for improved regulation and enforcement.[39] But arriving at a good remedial scheme for data-related wrongs requires taking seriously the significant social value at stake in the work product derived from data. This Article describes how a court or public agency considering model deletion should use traditional equitable factors to weigh the appropriateness of ordering the deletion of a trained model in the particular context of a lawsuit or enforcement action.[40]

[34] *See, e.g.*, Chopra, *supra* note 32, at 9 ("Corporate recidivism has become normalized and calculated as the cost of doing business; the result is a rinse-repeat cycle that dilutes legal standards . . . . Agency and court orders are not suggestions, but many large companies see them as such.").

[35] *See infra* Part II.

[36] *See* JEFFREY A. DEAN, DAEDALUS, A GOLDEN DECADE OF DEEP LEARNING: COMPUTING SYSTEMS & APPLICATIONS 62–66 (2022), https://www.amacad.org/publication/golden-decade-deep-learning-computing-systems-applications [https://perma.cc/R5SG-PW2T].

[37] *See, e.g.*, Omri Ben-Shahar, *Privacy Protection, At What Cost? Exploring the Regulatory Resistance to Data Technology in Auto Insurance*, 15 J. LEGAL ANALYSIS 129, 129 (2023).

[38] *See* NESTOR MASLEJ, LOREDANA FATTORINI, RAYMOND PERRAULT, YOLANDA GIL, VANESSA PARLI, NJENGA KARIUKI, EMILY CAPSTICK, ANKA REUEL, ERIK BRYNJOLFSSON, JOHN ETCHEMENDY, KATRINA LIGETT, TERAH LYONS, JAMES MANYIKA, JUAN CARLOS NIEBLES, YOAV SHOHAM, RUSSELL WALD, TOBI WALSH, ARMIN HAMRAH, LAPO SANTARLASCI, JULIA BETTS LOTUFO, ALEXANDRA ROME, ANDREW SHI & SUKRUT OAK, STANFORD UNIVERSITY, INSTITUTE FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, THE AI INDEX 2025 ANNUAL REPORT 246–59 (2025), https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf [https://perma.cc/TM68-EN9B] (providing estimates for the training costs for some of the largest generative AI models).

[39] *See, e.g.*, Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1355–69 (2023); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1374–76 (2020).

[40] *See infra* Part IV.

The Article proceeds as follows. Part I briefly describes the relationship between machine learning models and the data on which they train and then discusses the emergence of model deletion as a remedy over the last few years. Part II then examines the advantages of model deletion as a remedy. Next, Part III assesses the downsides of model deletion and, in particular, the possibility that it will be a severely disproportionate remedy in some scenarios. Finally, Part IV discusses how to manage the pluses and minuses of model deletion, building out a set of equitable factors that courts or other decisionmakers should consider when determining whether model deletion is appropriate.

I. MACHINE LEARNING MODELS AND THE EMERGENCE OF MODEL DELETION

What is model deletion, and where did it come from? To understand model deletion, it is necessary to first establish what is being deleted: the models developed from the data at issue in any particular case. This section provides a brief overview of the objects of model deletion orders—machine learning models—and how they relate to the data that is in dispute in these cases.[41] It then moves on to model deletion, in particular canvassing the cases in which it has been used or sought.

A. *Machine Learning Models*

In recent years, machine learning tools have become an increasingly significant part of commercial and social life and have drawn correspondingly increased attention from academics and policymakers.[42] "Machine learning" is a general term but refers broadly to the ability of computer programs to "learn" from data.[43] While machine learning is only one area of study within the field of artificial intelligence, machine learning tools are often colloquially referred to with the catch-all term "artificial intelligence" (or just "AI"), as they

[41] Because model deletion orders are phrased so broadly—encompassing all "work product" that is derived from a particular data set—it is possible that they could sweep up other kinds of work product aside from machine learning models. *See, e.g.*, Final Order at 4, *Cambridge Analytica, LLC*, *supra* note 8. But the focus of model deletion orders in practice, as well as the focus of the commentary surrounding model deletion orders, has been on machine learning tools. *See infra* Section I.B. This Article therefore discusses only machine learning applications of disputed data, rather than other types of work product that may be derived from data.

[42] *See, e.g.*, Lehr & Ohm, *supra* note 25, at 669–70 (discussing machine learning applications and associated legal scholarship).

[43] *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 651 (4th ed. 2021); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 89 (2014).

frequently take the form of computerized applications that seem "intelligent" in some form—such as making predictions, conducting analysis, or generating text.[44]

There are a variety of approaches to machine learning, but as a general matter, they all involve a process in which a computer analyzes data, creates a model based on that data, and uses that model in some way—to make predictions or otherwise solve problems.[45] The model is thus a core end product of the process, a tool embodying the learning that has occurred, and which can be used for a variety of purposes.[46]

This Article focuses on the type of machine learning that has accounted for much of the scientific and commercial growth in the field in recent decades: the use of sophisticated algorithms to "train" models on large data sets.[47] During the training process, data is run through an algorithm with the goal of identifying features of the data that the model can "learn" and apply in the future.[48] So, for instance, a machine learning model might be trained on images of handwritten numbers in order to "learn" how to recognize such numbers; when completed, it could then take an image of a handwritten number as an input and generate as an output its best guess as to what number it had just been shown.[49]

Machine learning models can be made in different ways, but a common underlying foundation for many of the successful tools created in recent years is a model architecture known as the multilayer neural network.[50] To oversimplify, a multilayer neural network is a type of model that is composed of layers of digital "simulated neurons," inspired in a rudimentary way by the structure of the human brain.[51] Like human neurons, these facsimile neurons

[44] *See* Surden, *supra* note 43, at 90 ("If performing well, machine learning algorithms may produce automated results that approximate those that would have been made by a similarly situated person.").

[45] RUSSELL & NORVIG, *supra* note 43, at 651.

[46] *See, e.g.*, *Model*, GENLAW: GLOSSARY, https://blog.genlaw.org/glossary.html [https://perma.cc/69JW-4MUF] ("Models are at the core of contemporary machine learning.").

[47] *See* RUSSELL & NORVIG, *supra* note 43, at 26, 750–51; *see also* Jeff Dean, David Patterson & Cliff Young, *A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution*, 32 IEEE MICRO 21 (2018).

[48] *See* Lehr & Ohm, *supra* note 25, at 695–701 (describing the model training process).

[49] MELANIE MITCHELL, ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS 36–37 (2019).

[50] *Id.* at 35.

[51] *Id.* at 22–29, 35–37 (using the phrase "simulated neurons" to describe the basic units of contemporary neural networks, and describing the history of perceptrons and their loose inspiration in human neurobiology).

(sometimes called "nodes") are connected to each other and pass information to each other along these connections.[52] As with a brain, much of the model's functionality resides in the strength and structure of these connections.[53] But unlike a brain, which adjusts the connections between neurons organically as it learns from experience, the strength of the connections between a model's nodes are learned during the model's training process.[54] As a model is trained, it assigns numbers called "parameters" or "weights" to the connections between its nodes, adjusting them to minimize its errors and improve its accuracy on the tasks (such as image recognition) given to it during training.[55] The weights that mediate the connections between neurons form the core of the model, and collectively determine what a model's output is for a given input.[56]

Training a model to obtain these weights can be expensive. The largest models may have hundreds of billions of weights, and training them can take significant resources of time, electricity, computing power, and data.[57] Recent reports and estimates put the cost of training the most powerful versions of large language models in the tens of millions or hundreds of millions of dollars.[58] The model weights that result from that training are best understood as the primary asset that is obtained by spending that money: the embodiment of the data, computing power, and electricity in a usable format that represents what the model has "learned" from the training process.[59] If you obtain the weights of a machine learning model, you can largely reproduce the model's capabilities without having to go through the training process, making the model weights a

[52] *Id.*

[53] *Id.* at 24.

[54] *Id.* at 26; *see also* RUSSELL & NORVIG, *supra* note 43, at 754–56 (describing core mechanics of the learning process in a simple neural network).

[55] *See* MITCHELL, *supra* note 49, at 36–38 (describing back-propagation); MICHAEL NIELSEN, *Using Neural Nets to Recognize Handwritten Digits*, *in* NEURAL NETWORKS AND DEEP LEARNING (2015), http://neuralnetworksanddeeplearning.com/chap1.html [https://perma.cc/RA99-3K68].

[56] Timothy B. Lee, *How Computers Got Shockingly Good at Recognizing Images*, ARS TECHNICA (Dec. 18, 2018), https://arstechnica.com/science/2018/12/how-computers-got-shockingly-good-at-recognizing-images/ [https://perma.cc/GKG6-NEQ9 (staff-uploaded archive)].

[57] *See* MASLEJ ET AL., *supra* note 38, at 49–51, 64–65.

[58] *See id.* at 64 (estimating that OpenAI's GPT-4 cost around $78 million and Google's Gemini Ultra cost around $191 million). Costs may change significantly over time. Anthropic's Claude 3.7 Sonnet, for instance, is a top-line model but reportedly cost "a few tens of millions of dollars" to train. Kyle Wiggers, *Anthropic's Latest Flagship AI Might Not Have Been Incredibly Costly to Train*, TECHCRUNCH (Feb. 25, 2025), https://techcrunch.com/2025/02/25/anthropics-latest-flagship-ai-might-not-have-been-incredibly-costly-to-train/ [https://perma.cc/X9JR-2E9H].

[59] *See, e.g.*, SELLA NEVO, DAN LAHAV, AJAY KARPUR, YOGEV BAR-ON, HENRY ALEXANDER BRADLEY & JEFF ALSTOTT, RAND, SECURING ARTIFICIAL INTELLIGENCE MODEL WEIGHTS: PREVENTING THEFT AND MISUSE OF FRONTIER MODELS 3 (Oct. 31, 2023).

core intellectual property and security interest of companies that spend significant funds on training proprietary models.[60]

Although the line between a trained model and its underlying training data is conceptually distinct, in practice, it can be blurry. In principle, the training data can be deleted, and the model could continue to be used, duplicated, licensed, or sold. But it is sometimes possible for a user to provide inputs to the model, intentionally or unintentionally, that prompt it to generate identical or near-identical copies of the data on which it was trained.[61] Especially in the context of commercial large language models, a model that outputs verbatim copies of its training data—which is called "regurgitation"—is often seen as undesirable for many reasons and a problem that model developers work to avoid.[62] Nonetheless, such regurgitation can happen, both as a matter of everyday use and especially as a result of actors intentionally trying to expose the underlying training data.[63] And one of the implications of such regurgitation is that parts of a model's training data can persist as complete or near-complete copies encoded in the model's parameters.[64]

Nonetheless, once trained, a model exists as a distinct tool that can use what it has "learned" in a variety of ways, depending on its design. Machine learning models can make predictions or inferences, such as estimating whether someone has or is likely to develop an illness.[65] They can generate content, such

[60] *Id.*; *see also* Sharon Goldman, *Why Anthropic and OpenAI Are Obsessed with Securing LLM Model Weights*, VENTUREBEAT (Dec. 15, 2023), https://venturebeat.com/ai/why-anthropic-and-openai-are-obsessed-with-securing-llm-model-weights/ [https://perma.cc/EM7G-WB4Q].

[61] *See, e.g.*, A. Feder Cooper & James Grimmelmann, *The Files Are in the Computer: Copyright, Memorization, and Generative-AI*, CHICAGO-KENT L. REV. (forthcoming 2025) (manuscript at 16–17)(discussing extraction and regurgitation). Some view this as contrary to the purpose of machine learning models, arguing that models are usually designed with the goal of formulating generalizable knowledge from training data rather than memorizing that data specifically. *Id.* at 52 (describing how AI companies "discuss memorization as a kind of 'bug'"); Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman & David Bau, *Erasing Concepts from Diffusion Models*, ARXIV 3 (June 21, 2023), https://arxiv.org/pdf/2303.07345 [https://perma.cc/JR2D-AVNP] ("While the traditional goal of machine learning is to generalize without memorization, large models are capable of exact memorization if specifically trained to do so, and unintentional memorization has also been observed in large-scale settings, including diffusion models.").

[62] Gandikota et al., *supra* note 61, at 3.

[63] *See, e.g.*, Cooper & Grimmelmann, *supra* note 61, at 46–50 (discussing regurgitation in the context of adversarial users).

[64] *Id.* at 19–27.

[65] *See, e.g.*, Ramachandran Rajalakshmi, Radhakrishnan Subashini, Ranjit Mohan Anjana & Viswanathan Mohan, *Automated Diabetic Retinopathy Detection in Smartphone-based Fundus Photography Using Artificial Intelligence*, 32 EYE 1138, 1138–39 (2018), https://www.nature.com/articles/s41433-018-

as a short text, an essay, a photograph, or artwork.[66] Improvements in our collective capacity to make these tools have led to their deployment across a wide range of applications, including: science, engineering, healthcare, education, government, and more.[67]

But while these tools have generated tremendous interest and widespread use, they have also brought concerns about harm and ensuing litigation.[68] For instance, the ability to make inferences from data combined with vast data sets that may contain a variety of nonpublic or difficult-to-access information creates major privacy issues.[69] Meanwhile, the creations of large generative AI models, along with their ability to generate copies (or near copies) and make texts or images using particular characters, motifs, or styles, has raised massive concerns about intellectual property.[70] The result has been a buildup in recent years of public enforcement actions and private litigation around the creation and use of AI tools.[71]

Many of the concerns surrounding these tools implicate questions about whether it is lawful for a particular developer of an AI tool to obtain or use the underlying training data that created the model.[72] It may be, for instance, that an AI tool was trained on data that was obtained in violation of a contractual agreement or a statutory prohibition.[73] Or it may be that the training of a model on copyrighted data itself violates the Copyright Act.[74] While there is not yet a significant amount of regulation or legislation focused on governing AI tools,

0064-9 [https://perma.cc/9P5Y-M6YB] (finding that an automated detection system was highly sensitive at detecting diabetic retinopathy).

[66] *See, e.g.*, Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1888–89 (2024) (discussing generative AI systems and depicting AI-generated artwork).

[67] DEAN ET AL., *supra* note 36, at 62–66.

[68] *See, e.g.*, Daniel Wilf-Townsend, *Artificial Intelligence and Aggregate Litigation*, 103 WASH. U. L. REV. (forthcoming 2026) (manuscript at 11–17).

[69] *See, e.g.*, Solow-Niederman, *Inference Economy*, *supra* note 7, at 388–95; Daniel J. Solove, *Artificial Intelligence and Privacy*, 77 FLA. L. REV. 1, 36–43 (2025).

[70] Sag, *supra* note 66, at 1890–94.

[71] *See* Wilf-Townsend, *supra* note 68, at 15–23 (discussing private litigation); *see infra* Section I.B. (discussing FTC enforcement actions).

[72] *See* Alicia Solow-Niederman, *Do Cases Generate Bad AI Law?*, 25 COLUM. SCI. & TECH. L. REV. 261, 265 & nn.11–12 (2024) [hereinafter Solow-Niederman, *Do Cases Generate Bad AI Law*] (describing and collecting cases).

[73] *See, e.g.*, *infra* Section I.B. (discussing the examples of Cambridge Analytica and Amazon Alexa).

[74] *See, e.g.*, Complaint ¶¶ 160–62, N.Y. Times v. OpenAI, 757 F. Supp. 3d 594 (S.D.N.Y. 2023) (No. 1:23 CV 11195) (making this assertion).

there is at least some law governing the acceptable use of certain kinds of data.[75] As a result, many of the enforcement actions and lawsuits around AI tools have focused on how a company obtained or used the training data deployed to build machine learning models.[76] It is in these contexts that questions of model deletion have arisen so far, as the next section goes on to describe.

B. *The Emergence of Model Deletion as a Remedy*

Model deletion was first used as a remedy in an FTC order resolving its investigation of Cambridge Analytica in 2019.[77] The FTC accused the company of unlawfully harvesting personal information from Facebook users, saying the company falsely represented that it did not collect identifiable information from those users.[78] The company then used that unlawfully obtained data to train algorithms to target political and commercial advertisements.[79] In its final order against the company, the FTC required not only that the company delete the personal information that it had acquired, but also that it "[d]elete or destroy . . . any information or work product, including any algorithms or equations, that originated, in whole or in part, from this" illicitly obtained data.[80]

The Cambridge Analytica order was followed by two similar orders, in investigations against the company Everalbum (doing business as Ever) and WW International (the company formerly known as Weight Watchers). The Everalbum case involved a photo storage application that trained a face recognition algorithm on its users' photos to enable its software to group photos together based on who they depicted.[81] The FTC alleged that Everalbum had unlawfully misrepresented its practices to its users, including misrepresenting their ability to control whether their photos were used for face recognition training and also deceiving users into wrongly believing that their photos were deleted when they deactivated their accounts.[82] In its final decision and order, the FTC required not only that Everalbum delete the information that its users thought had already been deleted, but also that it "delete or destroy any

[75] The United States lacks a comprehensive privacy law but does have a variety of sector-specific privacy laws, state laws, and consumer protection laws that can be applied in the privacy context. *See* Solow-Niederman, *Inference Economy*, *supra* note 7, at 368–78.

[76] *See infra* Section I.B.

[77] *See* Final Order at 3–4, *Cambridge Analytica, LLC*, *supra* note 8.

[78] *See* Complaint at 1, Cambridge Analytica, LLC, F.T.C. Docket No. 9383 (July 22, 2019).

[79] *Id.* at 1, 3.

[80] *See* Final Order at 4, *Cambridge Analytica, LLC*, *supra* note 8.

[81] Complaint at 1, Everalbum, Inc., F.T.C. Docket No. C-4743 (May 6, 2021).

[82] *Id.* at 6.

[a]ffected [w]ork [p]roduct," which it defined as "any models or algorithms developed in whole or in part using" the information in question.[83] Nearly identical language was included in the stipulated order in the WW International case, in which the FTC charged that the company violated the Children's Online Privacy Protection Act ("COPPA") by impermissibly gathering information about children and minors via its weight-loss app.[84]

Before these actions, orders of model deletion had not been used even in similar cases. In 2019, for instance, the FTC fined Google for harvesting children's data on YouTube without their parents' consent, in violation of COPPA and the FTC Act.[85] Commissioner Chopra dissented from the FTC's position, arguing that the fine was insufficiently punitive—and noting, among other things, that the "increased value of [YouTube's] predictive algorithm trained by ill-gotten data" would "not be reversed" by the monetary penalty.[86] Similarly, in an enforcement action against Facebook for violating an earlier FTC order limiting its use of facial recognition tools, the FTC had not sought to require Facebook to give up the benefits of technology derived from unlawfully obtained or used data.[87] Citing these decisions in the Everalbum case, Chopra referred to the new tack of ordering model deletion as "an important course correction."[88]

In addition to Commissioner Chopra, Commissioner Slaughter wrote publicly about these orders to specifically praise model deletion and note the novelty of the FTC's path. After the Cambridge Analytica, Everalbum, and Weight Watchers orders, Slaughter wrote in a co-authored article that

[83] Decision and Order at 2, 5, Everalbum, Inc., F.T.C. Docket No. C-4743 (May 6, 2021).

[84] *See* Stipulated Order for Permanent Injunction, Civil Penalty Judgment, and Other Relief at 2, United States v. Kurbo, Inc., No. 3:22 CV 946-TSH (N.D. Cal. Mar. 3, 2022) (defining "Affected Work Product"); *id.* at 8 (ordering the deletion of affected work product in language parallel to the complaint in the Everalbum case).

[85] *See* Stipulated Order for Permanent Injunction and Civil Penalty Judgment at 1, FTC v. Google LLC (D.D.C. Sept. 10, 2019) (No. 19 CV 02642).

[86] Dissenting Statement of Commissioner Rohit Chopra at 6, Google LLC and YouTube, LLC, F.T.C. File No. 1723083 (Sept. 4, 2019) [hereinafter Dissenting Statement Regarding Google and YouTube], https://www.ftc.gov/system/files/documents/public_statements/1542957/chopra_google_youtube_dissent.pdf [https://perma.cc/3722-2EBP].

[87] *See* Dissenting Statement of Commissioner Rohit Chopra at 1, Facebook, Inc., F.T.C. File No. 1823109 (July 24, 2019) [hereinafter Dissenting Statement Regarding Facebook], https://www.ftc.gov/system/files/documents/public_statements/1536911/chopra_dissenting_statement_on_facebook_7-24-19.pdf [https://perma.cc/4XCF-9GEY].

[88] Statement Regarding Everalbum, *supra* note 25, at 1.

algorithmic disgorgement was an "innovative remedy."[89] Discussing the Everalbum case, Slaughter said, "The premise is simple: when companies collect data illegally, they should not be able to profit from either the data or any algorithm developed using it."[90] She located the FTC's authority to seek the remedy in "the Commission's power to order relief reasonably tailored to the violation of the law" and said that the approach of model deletion "should send a clear message to companies engaging in illicit data collection in order to train AI models: Not worth it."[91]

A variety of outside commenters agreed. Professor Tiffany Li described this development as "one of the most revolutionary calls the FTC has ever made regarding AI."[92] Others praised model deletion as "an important legal remedy" that "levels the playing field for law enforcement";[93] a remedy that is "very welcome" to consumer advocates;[94] and a "compelling enforcement mechanism."[95] These commenters focused on model deletion's value as a robust deterrent, citing the significant costs that it might impose as a way of changing the incentives faced by businesses and limiting misconduct.[96] Some suggested that new legislation could create an individual right to model deletion as a "corrective remedy" for individual privacy violations.[97]

Although no law has gone that far, model deletion now appears to be an established and regular part of the FTC's toolkit. Since the wave of commentary

[89] Slaughter et al., *supra* note 10, at 39.

[90] *Id.*

[91] *Id.*

[92] Li, *supra* note 7, at 501.

[93] Hutson & Winters, *supra* note 10, at 151.

[94] Elder, *supra* note 10, at 1025.

[95] Eda Uludere, *Fruits of Deception: Model Destruction as an Enforcement Tool*, DATAETHICS (July 13, 2022), https://dataethics.eu/deceptive-data-practices-can-lead-to-ai-model-destruction/ [https://perma.cc/U3T3-QDQ5].

[96] *See, e.g.*, Li, *supra* note 7, at 503 ("Introducing algorithmic disgorgement as a privacy right in privacy law would increase the potential compliance burdens on companies but could also increase the deterrent effect, raising the risks to such an extent that companies would be encouraged to be even more careful with their use of data and machine learning."); Hutson & Winters, *supra* note 10, at 127–28 ("As a remedy and lever for law enforcement, model deletion would deter harmful AI and the broader framework of surveillance capitalism because it would disincentivize wanton data extractionism and incentivizes dataset accountability." (footnote omitted)); Elder, *supra* note 10, at 1024–25 ("[T]he FTC's decision to apply the IP deletion requirement liberally would broadly deter unlawful collection of consumer data, encouraging companies to follow both sector-specific privacy statutes and their own privacy policies.").

[97] *See* Brandon LaLonde, *Explaining Model Disgorgement*, IAPP (Dec. 13, 2023), https://iapp.org/news/a/explaining-model-disgorgement [https://perma.cc/UT8J-GHNR]; *see also* Li, *supra* note 7, at 504 (arguing for adopting model deletion as an individual privacy right).

after the initial three cases in which it was deployed, the FTC has continued to use the remedy. In 2023, for instance, when it charged the camera maker Ring with unlawful conduct around its treatment of customers' video recordings, the FTC required not only that the company delete the underlying recordings but also that it "delete or destroy any Affected Work Product unless such deletion is technically infeasible."[98] A similar order was entered that same year in its case against Edmodo, an educational technology provider that the FTC charged with violating COPPA.[99] And in 2024, when it charged Rite Aid with using a biased and unlawful face recognition system that misidentified customers as criminals, the FTC required not only the deletion of Rite Aid's photos and videos of customers that were used in the program but also "any data, models, or algorithms derived in whole or in part therefrom."[100]

Comparable language is present in orders in a variety of subsequent enforcement actions.[101] The FTC now often requires that companies delete work product derived from unlawfully obtained or possessed data. And in a recent Advanced Notice of Proposed Rulemaking on commercial surveillance and data security, the FTC indicated that it is considering formalizing this tool in regulation, identifying it broadly as "a remedy that forbids companies from profiting from unlawful practices related to their use of automated systems."[102]

That is not to say, though, that the FTC has used model deletion in every case in which it plausibly could. In 2023, for instance, the FTC entered an order against Amazon after charging that Amazon had violated COPPA and the FTC

[98] Stipulated Order for Injunction and Monetary Judgment at 7, FTC v. Ring LLC, No. 1:23 CV 1549 (D.D.C. June 16, 2023). As in prior actions, "Affected Work Product" included "any models or algorithms . . . developed in whole or in part from review and annotation" of the recordings in question. *Id.* at 2.

[99] Stipulated Order for Permanent Injunction and Civil Penalty Judgment at 3, 13, United States v. Edmodo, LLC, No. 23 CV 2495-TSH (N.D. Cal. June 27, 2023) (ordering the defendant to "delete or destroy any Affected Work Product," defined as "any models or algorithms developed in whole or in part using" the unlawfully obtained information at issue).

[100] Stipulated Order for Permanent Injunction and Relief at 13, FTC v. Rite Aid Corp., No. 2:23 CV 0523 (E.D. Pa. Feb. 26, 2024).

[101] *See* Decision and Order at 6, Avast Ltd., F.T.C. Docket No. C-4805 (June 26, 2024) (requiring the deletion of specific data and "any models, algorithms, or software developed . . . based on" that data); Decision at 3, 12, X-Mode Social, Inc., F.T.C. Docket No. C-4802 (Apr. 11, 2024) (requiring the deletion of all "Data Products," defined as "any model, algorithm or derived data . . . developed, in whole or part, using" specific data); Decision and Order, Mobilewalla, Inc., F.T.C. Docket No. C-4811 (Jan. 13, 2025) (using similar language); Decision and Order at 6, 11, Gravy Analytics, Inc., F.T.C. Docket No. C-4810, at 3, 12 (Jan. 13, 2025) (using similar language).

[102] *See* Trade Regulation Rule on Commercial Surveillance and Data Security, 87 Fed. Reg. 51273, 51285 (proposed Aug. 22, 2022) (to be codified at 16 C.F.R. ch. I).

Act with its Alexa device by recording and transcribing children's voices to train its voice recognition models to understand children and deceiving parents about deleting its recordings of their children.[103] But the FTC's remedial order did not require that Amazon delete any work product that it had already created with the impermissibly retained recordings.[104] Instead, it required that in the future, when recordings were deleted in accordance with the law, Amazon could not "subsequently use such information for the creation or improvement" of models or tools.[105] While some commenters have lumped this action in with discussions of model deletion,[106] it appears to be something distinct—a prospective limitation on the training of new models or the updating of old models, rather than a requirement that old models be destroyed.[107]

Model deletion has now begun to emerge outside of the context of the FTC as well. The State of Texas, for instance, has sought model deletion in an action against Meta for allegedly violating state biometric privacy laws.[108] But perhaps most significantly, private litigants have begun to seek model deletion as a remedy in civil litigation. One of the most important ongoing lawsuits in the world of artificial intelligence is the suit by the *New York Times* against the company OpenAI, in which the newspaper argues that the tech company's use of *Times* articles to train its large language models is a violation of copyright law.[109] In the prayer for relief in the *New York Times*' complaint, the newspaper seeks, among other things, "destruction . . . of all GPT or other LLM models and training sets that incorporate Times Works."[110] It grounds this request in the Copyright Act, which specifically provides that courts in copyright cases "may order the destruction or other reasonable disposition of all . . . articles by means of which" unlawful copies "may be reproduced."[111] And the *New York*

[103] *See* Complaint for Permanent Injunction, Civil Penalties, and Other Relief at 5–9, United States v. Amazon.com, Inc., No. 2:23 CV 00811-TL (W.D. Wash. May 31, 2023).

[104] Stipulated Order for Permanent Injunction, Civil Penalty Judgment, and Other Relief at 8, United States v. Amazon.com, Inc., No. 2:23 CV 00811-TL (W.D. Wash. July 19, 2023).

[105] *Id.*

[106] *See* Riley, *supra* note 14.

[107] Similarly, the FTC appears not to have sought model deletion in its case against video game maker Cognosphere, despite some of the allegations being that the company collected data unlawfully. *See* Stipulated Order for Permanent Injunction, Civil Penalty Judgment, and Other Relief at 15–18, United States v. Cognosphere, LLC, No. 2:25 CV 447 (C.D. Cal. Jan. 17, 2025).

[108] *See* Plaintiff's Petition at 25, Texas v. Meta Platforms, Inc., No. 22-0121 (71st Jud. Dist. Tex. July 30, 2024).

[109] *See* Complaint at 1–4, N.Y. Times Corp. v. OpenAI, 757 F. Supp. 3d 594 (S.D.N.Y. 2023) (No. 1:23 CV 11195).

[110] *Id.* at 68.

[111] *Id.*; 17 U.S.C. § 503(b).

Times is not the only one—model deletion has been requested in other lawsuits as well, with generative AI litigation still in its early days.[112]

Model deletion is thus both established and inchoate. The FTC has now been using it for years, and other public enforcers and private litigants are currently seeking it in major litigation efforts. But there is not yet any formal doctrine or guidance on model deletion—no court rulings, agency guidance, legislation, or regulations that specifically discuss the remedy or define the contexts in which it is appropriate. The next sections of this Article explore the advantages and disadvantages of model deletion, critiquing some of the existing portrayals of the remedy and working toward the contours of an equitable doctrine.

II. MODEL DELETION'S ADVANTAGES

The next two sections of this essay attempt to delineate some contours for a doctrine of model deletion. This section begins with model deletion's strengths as a remedy. The next section discusses model deletion's limitations.

To begin with, deleting the model at issue addresses harms that could be caused by the model's continued existence. An order that requires only the deletion of the underlying training data will limit ongoing harms associated with that data being unlawfully available. But just because the data ceases being available in the form that it was trained on does not mean that the harms associated with that data stop. Where the data has been used to train a model, the model has still "learned" from that data and can either disclose it or make inferences based on it. As Professor Li describes it, data used to train models leaves an "algorithmic shadow" in its wake, where even after it has been deleted in its original form it can have persistent effects via the model.[113]

A model trained on an individual's information, for instance, may still retain information about that individual even if the underlying training data has been deleted. Large language models have been prompted in ways that have caused them to output their training data, including individually identifiable

[112] *See* Samuelson, *supra* note 10 ("Four of the 16 generative AI copyright complaints explicitly ask courts to order generative AI defendants to destroy the models that were trained on their works. . . . Other generative AI copyright plaintiffs may eventually amend their complaints to ask for this remedy. Or they may ask for impoundment and destruction as part of a requested injunctive order.").

[113] Li, *supra* note 7, at 490.

information like names and addresses.[114] Sophisticated and intentional actors have at times been able to get significant amounts of information out of trained models, as when Google researchers extracted "over 10,000 unique verbatim-memorized training examples" from OpenAI's ChatGPT language model.[115] Where privacy or intellectual property concerns arise from the availability of the underlying training data, those concerns can persist when representations of that data can be generated by the model.[116]

Other harms may persist that are based not on the ability of a model to directly generate private data but instead on the model's ability to make inferences as a result of training on that private data. The privacy harms that can result from the use of large datasets often go beyond disclosure. Instead, they may involve the use of machine learning tools to classify people or make predictions about them: whether to sell them goods and services, treat them for medical conditions, or govern them.[117] As Professor Alicia Solow-Niederman writes, these kinds of predictions can be about information that "people might prefer not to disclose," such as sexual orientation, health status, race or ethnicity, or political or religious beliefs.[118] Because using these tools may involve making inferences about new persons, they can also result in harms that affect individuals whose information was not even in the underlying training data for the model at issue.[119] Information that is gathered in violation of contract terms or legal prohibitions can train machine learning tools that raise these concerns, and the tools' use can continue to cause privacy harms through their inferential capacities even if their training data has been destroyed.

In addition to missing these kinds of ongoing harms, injunctive relief that requires only data deletion (as opposed to model deletion) will likely be inadequate from a deterrence perspective. For a company that has wrongfully obtained or retained data, the goal of having that data is often going to be to *use*

[114] *See* Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr & Katherine Lee, *Scalable Extraction of Training Data from (Production) Language Models*, ARXIV 10 (Nov. 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2311.17035 [https://perma.cc/UV23-F486].

[115] *Id.* at 9.

[116] *See generally* Cooper & Grimmelmann, *supra* note 61 (discussing intellectual property concerns with memorized data in generative-AI systems).

[117] *See* Solow-Niederman, *Inference Economy*, *supra* note 10, at 384–88.

[118] *Id.*; *see also* Sandra Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioral Advertising*, 35 BERKELEY TECH. L.J. 367, 376–77 (2020) (noting that this type of information "can be inferred from online behavior without users ever being aware").

[119] *See* Solow-Niederman, *Inference Economy*, *supra* note 10, at 384–85; Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573, 610–11, 641–43 (2021).

the data, not simply to possess it. In the FTC actions described above, for instance, the companies involved sought to build tools and systems to help with some profit-making venture: selling advertisements, developing product features, improving security, and so on.[120] Allowing a company to retain its trained models might allow it to capture some (or maybe most) of the value of its wrongdoing. Ordering the deletion of the underlying data is still a meaningful penalty, as it deprives the company of resources for training further models. But even that penalty is lessened if a company can retain its existing models, as those existing models can help with future model-training efforts as well.[121] As a result, unless some other part of the remedial regime (such as fines or damages) is designed to be correspondingly more punitive, a regime that allows defendants to keep their models will inadequately deter defendants from engaging in the wrongdoing that created those models in the first place.[122]

But what about damages? It is possible for a damages regime to adequately deter misconduct at the same level that a model deletion remedy would. One can imagine, for instance, fines that are so large that a company faced with the choice would prefer to delete its model rather than pay—suggesting that a fine can be set high enough to provide comparable deterrence to model deletion. But there are a couple of limitations to the damages approach. First, in any currently existing damages regime, there is no particular reason to think that actual damages or statutory damages would approximate this level. And second, even where damages are set high enough from a deterrence perspective, there is still the problem of ongoing harm—if there is continued value to a company to use the model, and the use of the model (as opposed to the possession of the training data) does not itself result in a fine, the company will likely continue to use it.

In addition to its useful deterrent effect and ability to address ongoing harms, model deletion provides additional incidental benefits from the perspective of both public and private enforcement. As the previous paragraph suggests, from a public enforcement perspective, model deletion reduces the

[120] *See supra* Section I.B.

[121] *See, e.g.*, Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman & Jamie Shotton, *Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone*, *in* PROCEEDINGS OF THE IEEE/CVF INT'L CONF. ON COMPUT. VISION 3661, 3662–65 (2021) (describing how a model can be trained on synthetic images that were generated in part by models trained on real images).

[122] *See* Dissenting Statement Regarding Google and YouTube, *supra* note 86, at 6; *see also* Remarks of Samuel Levine, *supra* note 25, at 10–11, (describing model deletion as a way to "reverse structural incentives to maximize information collection and abuses").

need to engage in complex calculations about where to set fines to achieve optimal deterrence. Commissioner Chopra's dissent in the FTC's 2019 YouTube order is instructive as to how relying solely on fines can go awry.[123] It can be hard to calculate an appropriate level of financial penalty, both because of the difficulty of putting a dollar value on a harm and because of the challenge of assessing what fines are sufficient to deter large companies when the relevant misconduct is tied up in a lucrative business model.[124] If a model is taken to represent the value derived from an unlawful action, as public officials have repeatedly suggested,[125] then ordering the deletion of the model may avoid the need to try and assign a specific value to the benefit gained to a defendant by its wrongful conduct as might be required by a more traditional restitution or disgorgement remedy.[126]

Model deletion has a stronger set of benefits for private enforcement. In the realm of private enforcement, actions premised solely on retrospective damages often fall short when it comes to harms associated with data appropriation. This is most true in the privacy context, where courts often struggle to map privacy harms onto legal doctrines that are actionable and provide plaintiffs with damages and/or standing.[127] As Professor Ryan Calo has described, "[H]arm presents an especially acute challenge in the context of privacy."[128] For instance, in the run-of-the-mill context where an online retailer sells user data in breach of their user agreement or fails to take reasonable steps to keep data secure, there may in principle be liability, but there may be no compensatory damages, as courts are often hesitant to place a dollar value on the privacy loss associated with the wrong.[129] Statutory damages can help

[123] *See* Dissenting Statement Regarding Google and YouTube, *supra* note 86, at 6–7; *see also supra* notes 86–87 and accompanying text.

[124] *See* Dissenting Statement Regarding Google and YouTube, *supra* note 86, at 6–7.

[125] *See, e.g.*, Remarks of Samuel Levine, *supra* note 25, at 10–11 (describing model deletion as resulting from "the simple principle that companies should not be able to profit from illegal data practices"); Slaughter et al., *supra* note 10, at 39 (using similar language); Statement Regarding Everalbum, *supra* note 25, at 1 (describing an FTC model deletion order as "requir[ing] Everalbum to forfeit the fruits of its deception").

[126] Of course, this advantage depends on the accuracy of the assumption that the model represents the value derived from an unlawful action. As Part III discusses below, there are circumstances in which that will not be true because the unlawful data at issue accounts for only a small portion of the model's value.

[127] *See, e.g.*, Lauren Henry Scholz, *Privacy Remedies*, 94 IND. L.J. 653, 656–58 (2019).

[128] Ryan Calo, *Privacy Harm Exceptionalism*, 12 COLO. TECH. L.J. 361, 361 (2014).

[129] *See, e.g.*, Bernard Chao, *Privacy Losses as Wrongful Gains*, 106 IOWA L. REV. 555, 585 (2021) ("In the typical data breach case, personal information from millions of customers has been taken. There is undoubtedly real injury. The courts simply refuse to place a dollar value on that injury.").

address this problem somewhat, but Article III standing or its state-law analogues still can pose an obstacle.[130] As a result, traditional causes of action like tort and contract, and even causes of action under privacy-focused statutes, are often ineffective at policing privacy wrongs.[131]

Model deletion can at least partially mitigate these problems. It is a remedy, not a cause of action, so it cannot create a new basis for a lawsuit where none existed before. But model deletion as a remedy has features that may assist with obtaining standing and with addressing the blind spots of economic damages. First, model deletion is a prospective remedy, supporting standing for plaintiffs who have a reasonable substantive argument that the defendant's possession of a model poses ongoing harm, even if damages for past harms are uncertain or absent.[132] Second, model deletion is designed to address a wrongfully held asset of the defendant's and does not inherently need to be pegged to any sort of quantified harm to the plaintiff, mitigating the need to assign a dollar value to a privacy harm. Model deletion thus parallels (and likely would pair well with) the cause of action of unjust enrichment, which avoids some of the hurdles of privacy litigation by focusing on the defendant's gain rather than the plaintiff's loss.[133]

Model deletion thus has a variety of benefits associated with it, as its proponents argue. But it is not a flawless remedy. The next section addresses the downsides and limitations of model deletion, and Part IV explores where it is more likely or less likely to be appropriate.

III. THE LIMITATIONS OF MODEL DELETION

As the previous section described, model deletion has some significant benefits as a remedy. But the costs and limitations of model deletion have remained mostly unexplored.[134] This matters because, at least for now, the legal doctrines surrounding model deletion are general, discretionary and involve a

[130] *See id.* at 599–600.

[131] *Id.* at 558–59.

[132] *See, e.g.*, Campbell v. Facebook, Inc., 951 F.3d 1106, 1119–20 (9th Cir. 2020) (determining that Article III standing exists for plaintiffs seeking to enjoin ongoing conduct by Facebook that posed ongoing adverse privacy effects).

[133] *See, e.g.*, Chao, *supra* note 129, at 575 ("Because unjust enrichment does not focus on the plaintiff's injury but on the defendant's gain, unjust enrichment can step in and provide data loss victims a viable remedy."); *see also* Scholz, *supra* note 127, at 655 (describing restitution as "the quintessential privacy remedy").

[134] The most detailed examinations of the downsides associated with model deletion are in Li, *supra* note 7, at 504–05; Goland, *supra* note 10, at 43–47; and Hutson & Winters, *supra* note 10, at 144–48.

practical balancing of costs and benefits. And if any future policymaker sets out to develop new law or doctrine around model deletion, an account of the costs as well as the benefits will be useful for determining the content of that new law. This section therefore examines the costs and limitations of this remedy, while the next section examines factors that courts should consider when deciding whether to order it as a remedy.

A. *The Legal Basis for Model Deletion*

To begin with, it is important to note that the legal landscape on which model deletion exists is not particularly constraining. As of now, there are no statutes, regulations, or doctrines that directly control the question of when model deletion is appropriate. The FTC's authority to order model deletion has not been challenged so far, and as a result, the FTC has not had to articulate a specific source of legal authority for the remedy.[135] Commissioner Slaughter, writing in a law review article, has argued that model deletion can be supported under the FTC's section 5 authority to police unfair and deceptive acts and practices.[136] That authority provides a broad basis for substantive liability, speaking in general terms, and does not specify much as to remedies.[137] But both when it comes to the FTC's section 5 authority and other substantive provisions, the FTC is authorized to seek permanent injunctive relief under section 13(b) of the FTC Act.[138] And the test governing the propriety of that injunctive relief is similar to the general equitable test for an injunction, examining the balance of equities involved and requiring a showing of what is in the public interest.[139] The relevant inquiry for model deletion under this

[135] Goland, *supra* note 10, at 27. This is not particularly surprising. In the area of privacy law, where the FTC's model deletion activity has taken place, the FTC operates almost exclusively through settlement agreements—leaving very little case law articulating the contours of the FTC's legal authority. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 610–11 (2014).

[136] Slaughter et al., *supra* note 10, at 38–39; *see also* 15 U.S.C. § 45(a) (providing the FTC with the power to prohibit unfair or deceptive acts or practices).

[137] *See id.* § 45.

[138] *See id.* § 53(b).

[139] *Id.*; *see also* FTC v. World Travel Vacation Brokers, Inc., 861 F.2d 1020, 1024–32 (7th Cir. 1988) (discussing and applying the statutory language); *cf.* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388, 391 (2006) (outlining the traditional four-factor equitable test for permanent injunctive relief). Perhaps the most significant difference is that the FTC Act does not require the FTC to demonstrate that it has suffered an irreparable injury that cannot be adequately compensated by money damages. Notably, the Supreme Court recently interpreted Section 13(b) of the FTC Act as not authorizing equitable relief in the form of retrospective monetary relief, such as an order for monetary

standard, then, is essentially a policy inquiry—a consideration of the interests of the various parties involved as well as the public interest more broadly.

There is a similar approach in the copyright context that has been invoked in the ongoing litigation involving generative AI tools. Under the Copyright Act, a plaintiff (or the government) may seek not only the destruction of infringing copies of a work, but also of any "articles by means of which such copies . . . may be reproduced."[140] As discussed above, there are at least some circumstances in which a model may generate verbatim or near-verbatim copies of training data.[141] Where that makes it possible to use a model to generate copyrighted material, then, this statutory provision plausibly permits model deletion.

But in that scenario, model deletion would only be permitted, not mandated: The Copyright Act's language is permissive rather than mandatory, and courts evaluating whether to order deletion apply the familiar four-factor equitable test for permanent injunctive relief.[142] Under this test, a court assesses whether the plaintiff has suffered an irreparable injury; whether monetary damages are inadequate to compensate for that injury; whether the balance of hardships between the plaintiff and the defendant supports an injunction; and whether the interest of the broader public supports an injunction.[143]

The current law of model deletion is therefore largely open-ended. Both the FTC Act and the Copyright Act—the two statutes that have been invoked to support orders and requests for model deletion—provide little guidance around when model deletion is warranted, instead placing the decision within broad discretionary bounds that functionally amount to a basic policy judgment. And because the four-part test for injunctive relief is the broad default test for permanent injunctions, this kind of broad discretion is likely to be a feature of courts inquiring into model deletion under other statutes, too.

disgorgement or restitution. *See* AMG Cap. Mgmt., LLC v. FTC, 141 S. Ct. 1341, 1348 (2021). The Court was clear, though, that that provision does authorize forward looking, nonmonetary injunctive relief. *Id.* at 1349. Despite the use of the term "disgorgement" by some proponents of model deletion, the model deletion remedy is both nonmonetary and prospective, preventing the ongoing and future use of the model in question. Model deletion therefore should be unaffected by the ruling in *AMG Capital Management*. *See id.*; *see also* Hutson & Winters, *supra* note 10, at 136–37.

[140] 17 U.S.C. § 503(b).

[141] *See supra* Section I.A.

[142] *See, e.g.*, Softketeers, Inc. v. Regal W. Corp., No. 8:19-CV-00519-JWH, 2023 WL 2024701, at *8 (C.D. Cal. Feb. 7, 2023).

[143] *Id.* (citing Adobe Sys. Inc. v. Feather, 895 F. Supp. 2d 297, 303–04 (D. Conn. 2012)).

B. *The Potential Disproportionality of a "No Bad Bytes" Rule*

The primary problem with model deletion as it currently exists is that it can be a grossly disproportionate remedy. As it has been implemented so far, model deletion amounts to a "no bad bytes" rule: where a defendant has unlawfully obtained or retained data, model deletion seeks the deletion of algorithms or models derived "in whole or in part" using that data.[144] In other words, if any part of the model's training data violated the law, model deletion results in the deletion of that model regardless of the absolute amount of that data or the proportion of that data to the rest of the model's training data. A defendant may have a large quantity of unlawful data and have trained its model exclusively on that data; or it may have a small amount of unlawful data in a vast pool of lawful data that it has trained its model on. Either way, because the model was trained "in whole or in part" on unlawful data, model deletion as currently practiced will call for the deletion of the entire model.

This approach may be fine in some scenarios, including, arguably, some or all of the enforcement scenarios in which model deletion has been deployed so far. Although the FTC's complaints do not contain detailed information about the models or training processes involved, it seems reasonable to believe that in scenarios like the Cambridge Analytica or Everalbum enforcement actions, the models at issue were relatively specialized tools whose creation relied meaningfully on the unlawful data at issue. The unlawful acquired data in the Cambridge Analytica case, for instance, was data from tens of millions of Facebook users that Cambridge Analytica had invested heavily in obtaining with the apparent purpose of building the algorithmic targeting tools that would be the subject of the model deletion order.[145] In the Everalbum case, the defendant's misrepresentations about its use of users' data appear to have been made to all of its users outside of a few jurisdictions, and it used this data to train a special-purpose face recognition tool for its own product as well as enterprise customers.[146]

In both of these cases, the defendants' models appear to have derived their value in large part, or perhaps entirely, from unlawfully obtained data.[147] In

[144] *See supra* Section I.B.

[145] *See* Complaint, Cambridge Analytica, LLC, *supra* note 78, at 2–5.

[146] *See* Complaint, Everalbum, Inc., *supra* note 81, at 3.

[147] In the Everalbum case, the company had combined the face data from its users with other datasets. But these datasets were publicly available, suggesting that much of the commercial value of Everalbum's models derived from their proprietary data (which contained the unlawfully obtained user

these contexts, the analogy to the traditional remedy of disgorgement makes some sense. Where the goal is to have a remedy that is assessed in terms of the defendant's gain from wrongdoing, as is the case with disgorgement,[148] it makes sense to order the deletion of a model whose value is a somewhat reasonable approximation of the benefits associated with the defendant's wrongful conduct.[149]

But there will be many circumstances in which the deletion of an entire model is disproportionate to the defendant's wrongdoing. Where a defendant has trained a model on a large amount of data, and the unlawful data at issue is neither a significant portion of that data nor a distinctly valuable subset of that data, there may be no strong argument that a substantial fraction of the model's value derives from the unlawfully obtained data.[150] Especially if the cost of

images). *See, e.g.*, Thomas H. Davenport & Thomas C. Redman, Your Organization Needs a Proprietary Data Strategy, HARV. BUS. REV. (May 4, 2020), https://hbr.org/2020/05/your-organization-needs-a-proprietary-data-strategy [https://perma.cc/95ME-P37L] (describing how proprietary data is necessary for a competitive advantage in the commercial context of AI tools).

[148] *See, e.g.*, RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 51 (AM. L. INST. 2011) ("The object of restitution in such cases is to eliminate profit from wrongdoing while avoiding, so far as possible, the imposition of a penalty. Restitution remedies that pursue this object are often called 'disgorgement' or 'accounting.'").

[149] A full account of the value of the defendant's wrongful conduct would also need to address the value of the underlying data, which is only partially captured by accounting for the value of the trained model. A separate order to delete the underlying data is often justified, too, and has been pursued by the FTC as well. *See, e.g.*, Complaint, Cambridge Analytica, LLC, *supra* note 78, at 4.

[150] Establishing the value of a particular set of data's contribution to a model's training is a complex question with a variety of approaches. *See, e.g.*, Jinsung Yoon, Sercan O. Arik & Tomas Pfister, *Data Valuation Using Reinforcement Learning*, 119 PROC. MACH. LEARNING RSCH. 10824, 10824 (2020), https://proceedings.mlr.press/v119/yoon20a/yoon20a.pdf [https://perma.cc/U6QM-94T6]; Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song & Costas J. Spanos, *Towards Efficient Data Valuation Based on the Shapley Value*, 89 PROC. MACH. LEARNING RSCH. 1167, 1168 (2019), https://proceedings.mlr.press/v89/jia19a/jia19a.pdf [https://perma.cc/2SDK-W3NT]. But at least at a general level, two features of a dataset will tend to correspond to its value in training the model in question: the amount of the data and the quality of the data. *See, e.g.*, Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang & Yanqi Zhou, *Deep Learning Scaling is Predictable, Empirically*, ARXIV 13 (Dec. 1, 2017), https://arxiv.org/pdf/1712.00409 [https://perma.cc/Q3LE-KCML]; Venkat N. Gudivada, Amy Apon & Junhua Ding, *Data Quality Considerations for Big Data and Machine Learning*, 10 INT'L. J. ON ADVANCES SOFTWARE 1, 2 (2017). Quality, in turn, is a highly multifaceted and context dependent feature of data but can range from basic features of the data such as whether it is well labeled and accurate to more contingent features such as whether it is more relevant to a particular task at hand than other available data. *See, e.g.*, Boxin Zhao, Boxiang Lyu, Raul Castro Fernandez & Mladen Kolar, *Addressing Budget Allocation and Revenue Allocation in Data Market Environments Using an Adaptive Sampling Algorithm*, ARXIV 1 (June 5, 2023),

creating the model was high, the remedy of deleting the model in such contexts will be disproportionate to the value attributable to the wrongdoing at issue, and sometimes grossly so.[151]

This scenario is not just hypothetical—it is likely to be increasingly common, if not common already. Companies have begun training models such as large language models and image generation models with staggering amounts of data, including corpora of training data that amount to significant portions of all of the information available on the Internet.[152] Large language models' training data is now frequently measured in trillions of "tokens," units that correspond to words or fractions of words.[153] The phrase "it went swimmingly," for instance, would (for at least some models) be four tokens, with one token corresponding to "it," "went," "swimming," and "-ly."[154] OpenAI is reported to have trained its GPT-4 model on 13 trillion tokens.[155] For comparison, by one estimate a Google Books corpus of 40 million books contains about 4.8 trillion tokens.[156] The number of tokens used per word is different for different models,

https://arxiv.org/pdf/2306.02543 [https://perma.cc/LH86-S7K7] (describing potential features of a data market designed to allow data purchasers to find and obtain high-quality data, such as data that is particularly relevant to the tasks they are attempting to perform).

[151] Note that this analysis focuses on the value created by the use of unlawful data, rather than the harm caused by that use—that is in keeping with the general justification of model deletion by analogy to disgorgement, which focuses on the profits attributable to a wrongdoer's misconduct. A full weighing of the appropriateness of model deletion, though, would look to the harms that a plaintiff suffers in any particular case as well, as Section IV.B.2. discusses *infra*.

[152] *See, e.g.*, Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites That Make AI like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2024, 6:00 AM), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ [https://perma.cc/VF62-QNS7 (staff-uploaded, dark archive)].

[153] *See, e.g.*, Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain & Jianfeng Gao, *Large Language Models: A Survey*, ARXIV 3 (Feb. 20, 2024), https://arxiv.org/pdf/2402.06196v2 [https://perma.cc/HA2Q-7DFE]. For more about the concept of "tokens," see *What Are Tokens and How To Count Them?*, OPENAI, https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them [https://perma.cc/7ZNF-THPW (staff-uploaded archive)].

[154] *See Tokenizer*, OPENAI, https://platform.openai.com/tokenizer [https://perma.cc/N8FC-WAKW (staff-uploaded archive)].

[155] Minaee et al., *supra* note 153, at 4.

[156] Mark Cummins, *How Much LLM Training Data Is There, in the Limit?*, EDUCATING SILICON (May 9, 2024), https://www.educatingsilicon.com/2024/05/09/how-much-llm-training-data-is-there-in-the-limit/ [https://perma.cc/GJG4-Z9BP].

but by one approach the complete works of William Shakespeare amount to about 1,180,000 tokens, or about 0.000009% of the training data of GPT-4.[157]

The amount of training data used by each generation of tools is also growing rapidly, and at current growth rates, the training of advanced large language models would encompass essentially all publicly available text (all websites, books, academic articles, and so on) by the end of this decade.[158] Large language models trained by the largest companies in the industry represent one end of the spectrum. Some machine learning tools will train on much less data, potentially many orders of magnitude lower than the data large language models train on. But there are often returns to scale in training machine learning tools, and many important and common tools—not the least of which are LLMs—are trained on these extremely large corpora.[159]

In the context of these large models, small amounts of data may still be responsible for some of the value embodied in the model, but they often will not account for a significant fraction of that model's total value. Let's say, for instance, that the various versions of ChatGPT have trained on William Shakespeare's works—a plausible assumption, given that these works are in the public domain, frequently copied, and culturally central. Some users may find it useful for ChatGPT to be able to quote Shakespeare, mimic Shakespeare, or evaluate and critique an essay about Shakespeare—functionalities that would likely be improved by having trained on Shakespeare's works, along with the vast amount of commentary on Shakespeare that exists in other work that is likely within its training corpus.[160]

But much of the utility of that large language model comes from capabilities that likely would be impaired little or not at all if the model had not trained on Shakespeare's works: capabilities such as writing basic software code, building automated chatbots, facilitating document processing, or translating

[157] A rough rule of thumb is that 100 tokens equal about 75 words. *See What Are Tokens and How To Count Them?*, *supra* note 153. The complete works of William Shakespeare, meanwhile, amount to 884,647 words. *See Frequently Asked Questions About Shakespeare's Works*, FOLGER SHAKESPEARE LIBR., https://www.folger.edu/explore/shakespeares-works/frequently-asked-questions/ [https://perma.cc/V7U5-4JYU] (citing MARVIN SPEVACK, A COMPLETE AND SYSTEMATIC CONCORDANCE TO THE WORKS OF SHAKESPEARE (1968)). 884,647 words * 100 tokens / 75 words = 1,179,529 tokens. Then (1,179,529 / 13,000,000,000,000) * 100 = 0.00000907%.

[158] *See* Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim & Marius Hobbhahn, *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*, ARXIV 1 (June 4, 2024), https://arxiv.org/pdf/2211.04325 [https://perma.cc/N7N3-AVF5].

[159] *See, e.g.*, *id.*

[160] *See, e.g.*, Schaul et al., *supra* note 152 (describing material in the training corpus of a large language model).

materials from foreign languages.[161] Although they are relatively new, large language models' services have already generated revenues in the billions of dollars, a number that reflects a wide range of use cases across commercial sectors.[162] For a state-of-the-art large language model like GPT-4, the value of its services may be derived "in part" from Shakespeare's works, but its value is in no way reducible to that component of its data, nor do Shakespeare's works account for a large fraction of that value.

Although model deletion is often framed in terms of "disgorgement," neither the law nor the logic of disgorgement would support the remedy of model deletion in these scenarios—for instance, if it turned out to have been unlawful to include Shakespeare's works in GPT-4's training. Disgorgement, a remedy often associated with liabilities that sound in restitution or unjust enrichment, is built on the general principle that "[a] person is not permitted to profit by his own wrong."[163] Disgorgement can be attractive for a number of practical reasons, but one of the primary advantages is that it can provide for a remedy that is measured by a defendant's gain rather than a plaintiff's loss.[164]

Often, the defendant's gain and the plaintiff's loss will be similar—if a defendant steals a plaintiff's car, for instance, the market value of that car will likely be a reasonable approximation of both the defendant's gain and the plaintiff's loss. But in other scenarios, the plaintiff's loss may be smaller for a variety of reasons. In the privacy context, for instance, a defendant may be able to monetize a plaintiff's data in ways the plaintiff could not on their own. Or the harm to the plaintiff may not be due to the lost monetization value of the

[161] *See, e.g.*, Matt Renner & Matt A.V. Chaban, *601 Real-World Gen AI Use Cases from the World's Leading Organizations*, GOOGLE CLOUD, https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders [https://perma.cc/4BBQ-43RS] (last updated Apr. 9, 2025) (discussing various uses of Google's generative AI tool suite).

[162] *See Large Language Model (LLM) Markets 2024–2034 with OpenAI, Google, Meta, Microsoft, Tencent, & Yandex Set To Dominate the $85+ Billion Industry*, BUSINESS WIRE (June 7, 2024, 12:08 PM), https://www.businesswire.com/news/home/20240607372298/en/Large-Language-Model-LLM-Markets-2024-2034-with-OpenAI-Google-Meta-Microsoft-Tencent-Yandex-Set-to-Dominate-the-85-Billion-Industry---ResearchAndMarkets.com [https://perma.cc/22G8-E9GY (staff-uploaded archive)] (estimating the market for large language models' services at $6.4 billion in 2024).

[163] RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 3 (AM. L. INST. 2011). The concepts of disgorgement, unjust enrichment, and restitution are all interrelated and often confused with each other. *See id.* § 1 cmt. a (noting that the Restatement uses "the word 'restitution' to describe the cause of action as well as the remedy . . . despite the problems this usage creates"). Here, I attempt to follow the convention of using "disgorgement" to refer to a particular remedy measured by the value of a defendant's wrongdoing, using "unjust enrichment" to refer to a cause of action. *See, e.g.*, *id.* § 51(4). I largely do not discuss restitution as a distinct concept.

[164] *Id.* cmt. b.

data, but instead due to risks resulting from the potential disclosure or misuse of the data—and those risks may be difficult to quantify or may not legally be an adequate basis for recovery.[165] So, for instance, where a person's location data might be of some small monetary value to advertisers, the harm to the person of the disclosure of their location might be much more significant where it enables a stalker to find them.[166]

Disgorgement allows for a remedy pegged to the higher value. The plaintiff can choose: where their loss is higher than the defendant's gain, they can seek normal damages; but where the defendant's gain is higher than the plaintiff's loss, disgorgement is a useful tool. This aligns with normative intuitions that a wrongdoer should not be able to reap a windfall just because the victim's injury was lower than the wrongdoer's gain.[167] And, as a policy matter, it also helps promote effective deterrence where a bad actor would be able to profit even after paying compensatory damages.[168]

But an important feature of the doctrine of disgorgement is that the penalty a wrongdoer pays is measured by the specific profit attributable to their wrongful conduct. For disgorgement, the usual amount of the penalty "is the net profit attributable to the underlying wrong."[169] This feature of disgorgement is not ancillary. The question of how to determine what profit is attributable to a given wrong is often difficult, with a variety of doctrinal principles designed to address that question—and significant space taken up on the question in the Restatement of Restitution and Unjust Enrichment.[170] The test for attribution is not as simple as but-for causation, but to award profits in restitution, a court must be willing to conclude that the profits in question are not "unduly remote" from the conduct.[171] A claimant typically has the burden to provide evidence "permitting at least a reasonable approximation of the amount of the wrongful

[165] *See, e.g.*, Chao, *supra* note 129, at 576–77.

[166] *See, e.g.*, Thomas E. Kadri, *Brokered Abuse*, 3 J. FREE SPEECH L. 137, 141 (2023).

[167] *See, e.g.*, Triton Constr. Co. v. E. Shore Elec. Servs., Inc., No. 3290-VCP, 2009 WL 1387115, at *28 (Del. Ch. May 18, 2009), *aff'd*, 988 A.2d 938 (Del. 2010) ("Delaware law requires that improper gains . . . be recoverable by Triton even though no specific injury to Triton can be measured. Such a penalty . . . serves to discourage disloyalty and prevents an unjust windfall by stripping the profits gained from . . . disloyal acts.").

[168] Pamela Samuelson, John M. Golden & Mark P. Gergen, *Recalibrating the Disgorgement Remedy in Intellectual Property Cases*, 100 B.U. L. REV. 1999, 2029–30 (2020) (describing reasons why compensatory damages may undercompensate and under-deter).

[169] RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 51(4) (AM. L. INST. 2011).

[170] *See id.* § 42 cmt. h; *id.* § 51, cmts. e–i.

[171] *Id.* § 51(5)(a).

gain."[172] As the Massachusetts Supreme Court has put it, "[T]he over-all object is to render the ultimate recovery a sound reflection of the defendants' unjust enrichment due to the [wrongful conduct], and no more."[173] These principles stand squarely opposed to model deletion in cases involving disproportionate value.

Similarly, deleting models in these cases involving disproportionate value is not clearly supported by the rationales that FTC commissioners have offered in support of algorithmic disgorgement.[174] It may be that these officials do not use the word "disgorgement" out of a belief that algorithmic destruction is literally supported by the doctrine of disgorgement. They may instead use "disgorgement" as a kind of stand-in that means "a penalty that prevents a wrongdoer from benefitting from their wrongdoing." But even if that is the case, these rationales still suggest some sort of limitation of causation or attribution—that there be a connection between the use of unlawfully obtained training data and the value that a defendant derives from a model. And there is a gap between any penalty that is supposed to track the benefit conferred by possession of data and a remedy that requires the complete destruction of an asset derived only "in part" from that data. A "no bad bytes" rule will sweep up models that trained on any amount of unlawfully possessed data, no matter the degree of care used to avoid that result, the mental state of the model developer, or the profit actually derived from the data in question.

And although the language of disgorgement suggests a remedy more oriented toward unjust benefits, the problem of disproportionality also applies when it comes to harms. Unlike actual damages (or even statutory damages), model deletion as a remedy is not automatically pegged to some measurement or approximation of harm in the world, whether that be the number of times a violation has occurred, an assessment of monetary loss or pain and suffering, or another metric. Instead, model deletion presents an all-or-nothing penalty: if it applies, a model is destroyed in its entirety, regardless of the harm that has been caused or may be ongoing.

[172] *Id.* § 51(5)(d).

[173] USM Corp. v. Marson Fastener Corp., 467 N.E.2d 1271, 1276 (Mass. 1984) (internal quotation marks omitted).

[174] *See* Remarks of Samuel Levine, *supra* note 25, at 10–11; Slaughter et al., *supra* note 10, at 39–40; Statement Regarding Everalbum, *supra* note 25, at 1.

C. *Problems with Model Deletion's Disproportionality*

To some, model deletion's potential disproportionality may be a feature, not a bug. Commenters who like model deletion point to the size of large technology companies and the challenge of effectively deterring them from misconduct, suggesting that the potential high impact of model deletion could be useful.[175] If what you are looking for is a big stick, in other words, model deletion can fit that bill.

But there are reasons that the law often tries to achieve proportionality in its remedies, and they apply in the context of model deletion just as in other contexts. Most straightforwardly, when a remedy is too harsh, it may deter too much, chilling productive activity due to fears of a loss that is not proportioned to the harm caused.[176] OpenAI, for instance, has generated billions of dollars of economic activity.[177] But the current logic of model deletion would allow the destruction of its main assets—large language models—if it turns out those models were trained on one unlawfully used blog post of 500 words. In an economy governed by such a regime, it would not make sense to invest in the creation of these models in the first place, even if their existence would be a large net benefit to society.[178]

[175] *See, e.g.*, Kate Kaye, *The FTC's New Enforcement Weapon Spells Death for Algorithms*, PROTOCOL (Mar. 14, 2022), https://www.protocol.com/policy/ftc-algorithm-destroy-data-privacy [https://perma.cc/Z6AC-PQJ8?type=image] ("The Federal Trade Commission has struggled over the years to find ways to combat deceptive digital data practices using its limited set of enforcement options. Now it's landed on one that could have a big impact on tech companies: algorithmic destruction."); Hutson & Winters, *supra* note 10, at 150–51 (discussing the value of "strict and punitive enforcement," and arguing that model deletion "helps to break the asymmetry of information and power" between law enforcement and technology companies); Goland, *supra* note 10, at 39–40 ("[W]hile the FTC has levied massive fines on companies for deceptive practices connected to data collection, such fines generally proved to be ineffective in addressing privacy concerns.").

[176] *See* Li, *supra* note 7, at 504 (noting that model deletion "could harm smaller companies and potentially chill innovation"); *see also* A. Mitchell Polinsky & Steven Shavell, *Punitive Damages: An Economic Analysis*, 111 HARV. L. REV. 869, 879 (1998) (discussing "socially excessive precautions").

[177] Shirin Ghaffary, *OpenAI Doubles Annualized Revenue to $3.4 Billion, the Information Reports*, BLOOMBERG (June 12, 2024), https://www.bloomberg.com/news/articles/2024-06-12/openai-doubles-annualized-revenue-to-3-4-billion-information [https://perma.cc/HP35-BTJM (staff-uploaded, dark archive)].

[178] This argument does not depend on one believing that society is better off from OpenAI specifically or the current generation of large language models. It may be that, if models like OpenAI's have violated copyright with respect to huge portions of their training data, there is a justification to order their deletion (as the *New York Times* has requested in its lawsuit). The point here is just that model deletion in its current form makes no inquiry about the relative proportions of cost and value, allowing for situations in which even beneficial developments are chilled by disproportionate penalties.

Of course, there are justifications in some contexts for raising the cost of a fine or a penalty beyond simple equivalence to the harm caused by the relevant misconduct. For instance, where the costs of detection and enforcement mean that not all offenders will be caught, there is a good argument to increase penalties to adequately deter misconduct.[179] But even this approach must bear some relationship to the harm caused for the justification to work.[180] The "no bad bytes" approach to model deletion, which contains no considerations for proportionality, is easily susceptible to running awry and turning into overdeterrence. Even where one might think that a penalty should be some multiple of the harm created by a defendant's conduct, it is hard to see how that stance would support a penalty that amounted to *any* multiple of that harm.

Another potential problem is the incentives that a disproportionate penalty creates for litigation. If model deletion is on the table even in grossly disproportionate scenarios, that affects parties' incentives in a lawsuit and settlement negotiations. Consider, for instance, the strategic position in litigation of a company (such as OpenAI) whose existence and revenue depend more or less entirely on large models trained on a huge corpus of data. For such a company, an order requiring the deletion of their primary model or models could plausibly represent an existential threat. Under a "no bad bytes" approach to model deletion, any litigant with a claim about unlawful training data could invoke this threat, regardless of the actual harm that litigant has suffered. Such a scenario creates the possibility of "holdup" dynamics.[181] Any claimholder in this situation can functionally impose a tax on a model developer's continued existence—and a tax that bears no necessary connection to a theory of harm or compensatory relief.[182] Litigation thus seems less likely to result in just or fair outcomes and more likely to result in extortionate negotiations.

And there will be litigation. While it might be plausible that public enforcement agencies would exercise their discretion and seek model deletion only where it is more likely to be in the public interest, private litigants are

[179] *See, e.g.*, Polinsky & Shavell, *supra* note 176, at 889.

[180] *Id.* at 890 (noting the possibility of overdeterrence).

[181] *See* Mark A. Lemley & Philip J. Weiser, *Should Property or Liability Rules Govern Information?*, 85 TEX. L. REV. 783, 795–96 (2007) (describing holdup problems).

[182] *See, e.g.*, Christopher M. Newman, *Patent Infringement as Nuisance*, 59 CATH. U. L. REV. 61, 62 (2009) (discussing an analogous problem involving holdout dynamics "whenever a property owner's right to exclude gives him leverage over productive efforts whose value cannot be realized without making some use of that property"); Deepa Varadarajan, *Improvement Doctrines*, 21 GEO. MASON L. REV. 657, 676–77 (2014) (describing holdout dynamics in certain property conflicts where disproportionate remedies may be involved).

unlikely to care about the possibility of overdeterrence if they may be able to get a windfall.[183] Where the potential value of a litigant's claim is capped at an extremely high number, that high ceiling creates incentives both for litigants to gamble on less plausible claims and for litigants with more plausible claims to try and obtain higher-valued settlements. And where the value of such claims is pegged to the value of the asset held by the defendant, instead of the harm suffered by the plaintiff, there is no reason to think that the outcome of these lawsuits would correspond to any theory of distributive justice or social value.[184]

IV. TOWARD A MODEL DELETION TEST

As the last two sections described, model deletion leaves us with a dilemma. On the one hand, model deletion has some distinct advantages as a remedy that addresses ongoing harms and may provide more effective deterrence. But model deletion also has the possibility of being grossly disproportionate, resulting in unjust and counterproductive outcomes. So how is a decision-maker—whether an enforcement agency or a judge overseeing private litigation—supposed to decide when model deletion is appropriate?

This section outlines the contours of a test for determining whether to use model deletion in a given case. Because the law governing model deletion is open-ended and mostly grounded in equity,[185] this test is guided primarily by traditional equitable principles. This section is written with a focus on judicial actors but could just as well apply to other enforcers such as federal agencies, and several considerations that depend on the identity of the enforcer are flagged at different points.

A. *Proportionality*

Imagine you are a judge considering the appropriate remedy in a case involving a machine learning model. The plaintiff has established that the defendant broke the law by obtaining or possessing a particular set of data. The plaintiff has also established that the defendant trained a machine learning model on that data. Now the plaintiff asks you to order that the defendant delete

[183] *See, e.g.*, Richard A. Bierschbach & Alex Stein, *Overenforcement*, 93 GEO. L.J. 1743, 1777 (2005) ("A private plaintiff's overarching criterion for filing a lawsuit is the difference between the expected investment in and the expected return from the litigation. She is not concerned about overenforcement and whether refraining from suit would help counteract its effects.").

[184] Newman, *supra* note 182, at 63. To the contrary, it may be more likely that distributive justice weighs in favor of allowing the defendant to recoup more of the gain from their investment, rather than providing the plaintiff with what can amount to a disproportionate windfall. *Id.*

[185] *See supra* Section III.A.

the model. The law is open-ended and structured primarily by traditional equitable principles, which appeal largely to abstract concepts such as fairness, culpability, and even-handedness. What should you do, given this flexibility and the facts in front of you?

Given that the primary problem with model deletion is its capacity to be a disproportionate penalty,[186] a good place to start would be to establish a sense of the proportions involved here. In other words, the first step of the process should be to at least get a rough sense of (a) the value of the model involved, and (b) how much of that value is derived from the unlawful data at issue. This may be difficult to pin down precisely, but it may not be necessary to get a very precise estimation of either (a) or (b) to get a reasonable and actionable sense of whether the unlawful data is responsible for a large or small fraction of the value of the model.

For instance, it may be hard to know precisely what the value of OpenAI's GPT-4 is,[187] and it may be difficult or impossible to get a precise quantification of how much of that value is derived from, e.g., the 1985 novel *The Cider House Rules* by John Irving (assuming that the novel was in GPT-4's training data). But it may be relatively straightforward to establish with adequate confidence that that novel is not responsible for a significant fraction of the model's value. Similarly, where a company appears to have relied significantly or exclusively on unlawful data to train its model—as seems likely in, e.g., the FTC's Cambridge Analytica case[188]—it would be unnecessary to pin down precise values to be adequately confident that the unlawful data is a major (or the sole) source of the model's value.

As these examples suggest, where models are large and trained on vast arrays of data or where they are narrower and trained on more specialized data, the proportionality inquiry will likely be easier. Those two types of models may account for many of the commercial use cases that get developed.[189] But there

[186] *See supra* Section III.B.

[187] Goldman Sachs, for instance, recently released a research report suggesting that investors may have overvalued AI tools and the companies that produce them in recent rounds of investments. *See* Allison Nathan, Jenny Grimberg & Ashley Rhodes, *Top of Mind: Gen AI: Too Much Spend, Too Little Benefit?*, GOLDMAN SACHS (June 25, 2024, 5:10 PM), https://www.goldmansachs.com/intelligence/pages/gs-research/gen-ai-too-much-spend-too-little-benefit/report.pdf [https://perma.cc/G56X-Q2BY (staff-uploaded archive)].

[188] *See supra* notes 77–80 and accompanying text.

[189] *See, e.g.*, LAREINA YEE, MICHAEL CHUI, ROGER ROBERTS & MENA ISSLER, MCKINSEY DIGITAL, TECHNOLOGY TRENDS OUTLOOK 2024 14–21 (2024) (describing trends involving both large general-purpose foundation models and smaller models trained on private, proprietary, and company-specific data).

are also many variations possible. For example, it is possible for an actor to take a broad, preexisting model as a foundation and fine tune it with a narrower, more specialized type of data to make the model fit for a particular purpose.[190] These circumstances will raise a variety of questions: if that specialized data was the unlawful data at issue, how do you evaluate its contribution to the ultimate post-fine-tuning model as compared with the data used for the broad, preexisting model? That kind of question will need care and attention. But difficult valuation arises in many remedial contexts—not the least of which is traditional disgorgement—and courts will be able to develop tools and doctrines to assess these kinds of questions in an appropriate, context-sensitive way.[191]

Depending on the outcome of the proportionality inquiry, it may be all that is necessary. If the unlawful data is responsible for the lion's share of the value of the model, model deletion would likely be an appropriate remedy: it passes the proportionality test. The concerns about model deletion being a disproportionate remedy are much weaker in this scenario, while the advantages of model deletion remain strong.

One disadvantage of model deletion would remain, though, that is worth keeping in mind in public enforcement contexts in particular: model deletion's destruction of something valuable rather than the transfer of that value. In a private lawsuit that meets the proportionality test, this is not a huge problem because the parties can bargain with each other over the model deletion order itself. If the model's continued existence is worth more than however much the plaintiff values its destruction, there should be a bargaining opportunity—the defendant can pay the plaintiff to sign over its right to enforce the order.[192] The plaintiff will be in a strong bargaining position and able to capture a very high proportion of the value of the model; but this seems reasonable, given that the

[190] *See, e.g.*, Thibault Schrepel & Alex 'Sandy' Pentland, *Competition Between AI Foundation Models: Dynamics and Policy Recommendations* 4–6 (MIT Connection Sci., Working Paper No. 1-2023, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4493900 [https://perma.cc/MV69-XKK9 (staff-uploaded archive)].

[191] *See* RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 51 cmt. e (AM. L. INST. 2011) (noting the complexity of disgorgement calculations and the use of presumptions to manage them); *see also* SEC v. Platforms Wireless Int'l Corp., 617 F.3d 1072, 1096 (9th Cir. 2010) (discussing and applying a "reasonable approximation" standard).

[192] *Cf.* Louise A. Halper, Nuisance, Courts and Markets in the New York Court of Appeals, 1850-1915, 54 ALB. L. REV. 301, 352–53 (1990); *see also* Mark P. Gergen, John M. Golden & Henry E. Smith, The Supreme Court's Accidental Revolution? The Test for Permanent Injunctions, 112 COLUM. L. REV. 203, 239 n.125 (2012) ("[T]he practical effect of the injunction in Whalen was to compel the mill to buy out the plaintiffs under the eye of the trial judge responsible for administering the injunction.").

model owes most of its value to a violation of some legal entitlement of the plaintiff's.

This logic is less possible in a world of public enforcement, though. Public enforcers may not see such a negotiation as a legitimate use of their authority—as it amounts to a defendant giving the enforcer a revenue stream in exchange for permission to continue benefitting from lawbreaking activity. And even if the enforcer did see such a negotiation as hypothetically worthwhile, it is difficult for the enforcer to know the subjective valuation of the relevant victims involved—which is a key part of assessing whether allowing the model to continue to exist is "worth it." In the private litigation scenario, the plaintiff internalizes the cost of the model's continued existence. But there is not a similar mechanism for a public enforcer. As a result, even where the proportionality test is passed, public enforcers may want to take some extra care to ensure that model deletion is not unproductively destroying something valuable, as there is not the backstop of private negotiation to safeguard against a deletion order that is excessive in relation to whatever ongoing problem is caused by the model's existence.

B. *Equitable Factors Favoring Deletion*

A more difficult situation arises where the unlawful data at issue is not responsible for most of the value of the trained model. Here, model deletion's disproportionality looms large and may make it an inappropriate remedy. But before ruling it out, a court should consider whether there are additional factors that may permit model deletion even if it is disproportionate when viewed solely through the lens of monetary valuation. Because the law supporting model deletion is fundamentally equitable, this section relies on traditional equitable factors—the defendant's culpability and the balance of hardships between parties—that may be relevant to the model deletion inquiry.

1. Culpability

One equitable factor that a court should examine when model deletion is being considered is the defendant's degree of culpability—for instance, if the defendant willfully violated the law, engaged in sharp practices, or otherwise had some sort of unclean hands. These kinds of considerations have traditionally justified imposing a harsher penalty or refusing to grant a benefit in equitable contexts.[193] Statutory penalties, too, often impose harsher penalties for willful conduct.[194] Such an approach has a policy rationale: the marginal increase in penalty provides more incentive to avoid breaking the law to those who are aware that they may be about to break the law, while avoiding being overly harsh to more "innocent," unknowing violators.[195] Such an approach also helps avoid chilling socially productive behavior by lowering the risk of a significant penalty for those who are engaging in conduct that seems ex ante to have a low risk of being unlawful.[196]

In the contexts in which model deletion is likely to arise, considerations of culpability may play a significant role in differentiating between when it is and is not acceptable. As the FTC's enforcement actions have indicated, there have been routine and repeat violations of privacy laws by large technology companies of a type that could support a finding of willfulness or conscious wrongdoing.[197] Facebook, for instance, engaged in repeated misconduct over years, including when it was subject to an FTC order—such as allowing Cambridge Analytica to access tens of millions of consumers' data despite promises to keep individual information private unless there was explicit consent to share it.[198] In the matter of Amazon's harvesting of data via its Alexa app, the FTC noted that Amazon repeatedly "discovered," many times over several years, that it had been failing to delete location information despite customers' deletion requests—but each time failed to correct the problem.[199] These sorts of repeated, knowing violations are the type of evidence that

[193] *See, e.g.*, Kansas v. Nebraska, 574 U.S. 445, 463 (2015) (pointing to how "Nebraska took full advantage of its favorable position" and the need to "deter[] future violations" as part of the justification for providing a remedy greater than actual damages); RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 63 (AM. L. INST. 2011) (limiting the availability of restitution where a claimant has engaged in inequitable conduct); *see also* T. Leigh Anenson, *Announcing the "Clean Hands" Doctrine*, 51 U.C. DAVIS L. REV. 1827, 1840–47 (2018) (describing the moral norms underpinning equity doctrines).

[194] *Compare* 15 U.S.C. § 1681n *with* § 1681o (providing for additional penalties for willful noncompliance compared with negligent noncompliance).

[195] *See, e.g.*, Samuelson et al., *supra* note 168, at 2077 ("Parties who are conscious of a specific and substantial risk of infringement are the ones whom the law can most productively encourage to seek out right holders and to negotiate for the right to use another's IP.").

[196] *Id.*

[197] *See, e.g.*, Rohit Chopra, *Lessons from the FTC's Facebook Saga*, REGUL. REV. (Sept. 27, 2022), https://www.theregreview.org/2022/09/27/chopra-lessons-from-the-ftcs-facebook-saga/ [https://perma.cc/4XMW-GAZ7].

[198] *Id.*

[199] Complaint at 9, United States v. Amazon.com, No. 23 CV 00811 (W.D. Wash. May 31, 2023).

traditionally can support a finding of willfulness.[200] In contexts where repeated disregard for legal obligations is part of the course of conduct giving rise to the model at issue, model deletion may be justified where it would not otherwise be because of concerns of disproportionality.

Conversely, a lack of culpability would weigh against model deletion. Scenarios involving either innocent or merely negligent defendants are also relatively easy to imagine in contexts where model deletion is at issue. The supply chains for machine learning models, and generative AI models in particular, are often long, complex, and opaque.[201] The datasets used to train models are often created and curated by actors other than those who create the models.[202] The quantity of data required to train large models often makes it infeasible to determine the origins of every item in a dataset, and those who provide datasets often provide imperfect or flawed accounts of what data they contain.[203] And there is also a good deal of legal uncertainty regarding how law applies in the context of machine learning models.[204] This all combines to raise the possibility of actors who use trained models with a good faith belief in the legality of their actions and the legal status of the underlying data, but who end up being liable later on—whether due to factual errors or misunderstanding regarding the data's provenance, or a new legal interpretation that resolves considerable uncertainty in a way that cuts against them. In such circumstances, evidence of the actor's care and good faith would tend to weigh against model deletion, where that model deletion would entail destruction disproportionate to the actor's wrongdoing.

2. The Balance of Hardships

In keeping with the framing of model deletion as loosely tied to the concept of disgorgement, the factors that have been considered so far focus on the defendant: the value of the defendant's model and how much is attributable to the defendant's wrongful conduct, and whether the defendant's culpability

[200] *See, e.g.*, Firestorm Pyrotechnics, Inc. v. Dettelbach, 61 F.4th 768, 775–76 (10th Cir. 2023); *Willfulness*, BLACK'S LAW DICTIONARY (12th ed. 2024) (defining willfulness to include "[t]he voluntary, intentional violation or disregard of a known legal duty").

[201] *See* Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, 72 J. COPYRIGHT SOC'Y 251 (2025) (manuscript at 307–47).

[202] *Id.* at 314–18.

[203] *Id.*

[204] *See, e.g.*, Sag, *supra* note 66, at 1890 ("The rise of generative AI poses important questions for copyright law.").

supports a disproportionate penalty. But an assessment of the propriety of model deletion should examine the harm caused by the defendant's conduct as well. In contrast to disgorgement, which typically involves an exchange of money, model deletion results in the destruction of an asset—an irrevocable and potentially extreme loss. And a primary justification for that deletion is the prevention of an ongoing wrong, rather than the remedying of a past wrong.[205] Considering the strength of that justification, then, requires an assessment of the kind and degree of harm that would be involved in allowing the model to persist.

A harm inquiry is also supported by traditional equitable considerations. First, equity typically directs courts to consider whether the harm at issue is ongoing, and whether it can be adequately addressed via money damages.[206] An order to delete a model is a permanent injunction, and permanent injunctions in many contexts are available only to prevent harms that cannot be remedied through damages.[207] This may not be as imposing a hurdle as it seems at first. Even where ongoing harm could conceivably be remedied with damages, the need to continuously monitor and enforce the legal entitlement to those damages can justify an injunction, for instance.[208] But asking whether a deletion order is actually necessary to address an ongoing harm is a reasonable step to take, particularly where deletion would result in a disproportionate penalty.

The second, and more significant, part of the harm inquiry should look to the balance of hardships. When courts issue injunctions, a regular part of the inquiry is a consideration of the relative hardship to the defendant if the injunction is granted and the plaintiff if the injunction is denied.[209] Injunctions, of course, are going to impose a burden on defendants, and courts reject the idea that they are engaged in simply balancing the plaintiff's interests against those of the defendant.[210] But where the harm to the plaintiff is noticeably less

[205] *See, e.g.*, Li, *supra* note 7, at 502.

[206] *See* eBay, Inc. v. MercExchange, L.L.C., 547 U.S. 388, 391 (2006) ("[A] plaintiff seeking a permanent injunction must . . . demonstrate . . . that remedies available at law, such as monetary damages, are inadequate to compensate for that injury.").

[207] *Id.* Notably, one context where irreparable harm may not need to be proved is public enforcement—for instance, the FTC is not required to prove irreparable harm to seek injunctive relief. *See supra* note 139.

[208] *See, e.g.*, Softketeers, Inc. v. Regal W. Corp., No. 8:19-CV-00519-JWH, 2023 WL 2024701, at *10 (C.D. Cal. Feb. 7, 2023).

[209] *See, e.g.*, *eBay*, 547 U.S. at 391; Restatement (Second) of Torts § 941 (Am. L. Inst. 1979).

[210] *See, e.g.*, Golden Press, Inc. v. Rylands, 235 P.2d 592, 595 (Colo. 1951) ("While the mere balance of convenience is not the proper test, yet relative hardship may properly be considered and the court should not become a party to extortion.").

serious than the harm to the defendant, the appropriate path for a court weighing the equities is to deny the injunction in question and leave the plaintiff to recover damages only.[211]

One illustration with some potential parallels to model deletion comes from the case law surrounding encroaching structures.[212] In these cases, a landowner builds a large structure on what they think is their land, but it turns out the structure extends somewhat onto adjacent land owned by someone else.[213] In these circumstances, courts examine the relative hardships to the parties of requiring the encroachment to be removed (which may result in tearing down or damaging the building) versus allowing it to stand (which may entail a forced sale of the portion of the land on which the encroachment stands to the landowner who built the building).[214]

Although the substantive provisions of trespass law typically mandate removal of the structure, courts have often held that equitable considerations make such an injunction discretionary.[215] And courts have usually declined to grant such an injunction where the intrusion on property is small; the cost of getting rid of the intrusion would be high; and the builder of the building did not know that it was a trespass.[216]

A similar doctrine, the doctrine of accession, exists in the world of chattels. Under that doctrine, where someone mistakenly takes another person's chattels and improves them—for instance, mistakenly taking another's lumber and building something out of it—the balance of the equities often weighs in favor of allowing the improver to keep the improved item. The improver is required to pay as damages only the price of the raw materials to the original owner.[217] These doctrines effectively convert property rules to liability rules in situations where there are grossly disproportionate values at stake and an innocent actor.

Some model deletion scenarios may parallel these cases. Consider, for instance, the creation of a large and expensive model that involves training on significant amounts of data curated by third parties. There is a rough parallel to

[211] *Id.*

[212] *See, e.g.*, *id.* at 595.

[213] *See, e.g.*, *id.*; Graham v. Jules Inv., Inc., 356 P.3d 986, 990 (Colo. App. 2014) (collecting cases); Soma v. Zurawski, 772 N.W.2d 724, 726–27 (Wis. Ct. App. 2009); Szymczak v. LaFerrara, 655 A.2d 76, 81 (N.J. Super. Ct. App. Div. 1995).

[214] *See, e.g.*, Hirshfield v. Schwartz, 110 Cal. Rptr. 2d 861, 866–67 (Cal. Ct. App. 2001).

[215] David A. Dana & Nadav Shoked, *Property's Edges*, 60 B.C. L. Rev. 753, 772 (2019).

[216] *Id.*

[217] *See* Varadarajan, *supra* note 182, at 667–68; *see also* Gergen et al., *supra* note 192, at 248–49 (noting the similarity between accession and the building encroachment cases).

the encroaching structures cases where it turns out that (a) a small, not particularly valuable subset of that data was unlawfully obtained or retained; (b) the defendant did not know about the existence of that data, had no reason to doubt its legality, or relied on affirmative representations of its legality; (c) it would be extremely expensive to retrain the model without that data; and (d) the continued existence of the model is not expected to cause much harm to the plaintiff. In these circumstances, the balance of hardships provides a strong argument against model deletion.

Of course, it may also be that the balance of hardships favors the plaintiff. In particular, (a), (b), and (c) in the previous paragraph can all be true in scenarios where (d) is false, and the continued existence of the model poses a real ongoing threat or guarantee of harm to the plaintiff. Such a scenario might arise where the model is trained on private information about an individual, and there is no way to prevent users of the model from accessing or making inferences based on that information. Or it might arise where the model is trained on copyrighted data, and there is no way to prevent people from using the model to generate unlawful copies. Harms in realms like privacy and intellectual property can be particularly abstract, so courts will need to be empathetic and conscientious to plaintiffs when weighing the balance of hardships in situations involving ongoing harm.[218]

It is also important to recognize that, although courts and commentators talk about the "balance" of the equities, there is a thumb on the scale against the defendant—who, after all, has violated the law.[219] In the encroaching structures cases, for instance, courts have called an injunction against the trespass as the "preferred equitable remedy," and often end up ordering the structure at issue to be removed, altered, or destroyed where the balance of hardships does not provide an adequate defense to such an injunction.[220] Model deletion orders

[218] Another likely possibility is a scenario where the developer did have reason to believe that its possession or use of the data in question was wrongful in some way. In that situation, another plausible analogue would be the "wrongful improvers" doctrine, which is like the accession or encroachment cases, but which involves a knowing wrongdoer. *See* Elder, *supra* note 10, at 1009. In those cases, courts are much more willing to enjoin the defendant and award title over the property in question to the plaintiff. *Id.* at 1032–33.

[219] *See, e.g.*, Varadarajan, *supra* note 182, at 682 (discussing how "a showing of minor or marginal improvement is generally insufficient to convince courts to depart from property rule protection," and instead "courts insist on a significant disparity of value" to rule for a defendant in cases such as accession).

[220] *See, e.g.*, Hunter v. Mansell, 240 P.3d 469, 479 (Colo. App. 2010) ("The overwhelming majority of case law in this jurisdiction demonstrates that the traditional and preferred equitable

may be similarly justified by ongoing harms caused by trained models that violate privacy interests, intellectual property interests, or other legally protected interests.[221]

Finally, when assessing the hardships involved in the potential exercise of model deletion, a court should consider the interests of third parties as well. Courts weighing injunctive relief are generally directed to consider the general public's interest.[222] In the model deletion context, that consideration could point in either direction. A defendant's model may be a source of widespread privacy violations or intellectual property violations, and ordering its destruction may benefit many beyond the plaintiff in the case at hand. Or a model may be incorporated into the operations of many third-party businesses, and its destruction could have negative impacts beyond the defendant's bottom line. Or there may be a strong noncommercial public interest in not having the model destroyed, such as if it has scientific or medical value.[223] The public interest in any given case may not be clear or decisive; but equity directs courts to consider interests beyond the parties to the case, and model deletion orders could easily have positive or negative spillover effects.[224]

remedy for a continuing trespass is a mandatory injunction requiring the removal of the encroachment. In our view, *Golden Press* does not change that."). The tendency to only deny injunctions where there is disproportionate harm is why this section of the paper places the discussion of undue hardship after the proportionality determination above in subsection A. Where a model deletion order would not result in disproportionate harm, the balance of the equities is unlikely to weigh against the order.

[221] A special concern arises where individual harms may be small but there are a large number of individuals who have been harmed. It is easy to think of such scenarios: massive state-of-the-art large language models cost hundreds of millions of dollars to train, and the balance of the equities seems to favor their owners when going against, for instance, the copyright interests of a person who wrote a handful of online posts or essays that became part of the model's training data. But the aggregate interests of all such persons may in fact be enough to weigh significantly in a court's calculus. The best way to manage this concern is likely through aggregate litigation, which has safeguards to foster adequate representation of the many persons whose interests would be implicated by such a case. *See* Wilf-Townsend, *supra* note 68, at 1 (arguing that aggregate litigation provides a number of advantages for resolving liability and remedial issues with AI tools). But in considering the balance of hardships and the equities, a court may want to consider whether the model's continued existence poses harms not just to the plaintiff, but to others as well.

[222] *See, e.g.*, eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388, 391 (2006).

[223] *See, e.g.*, Jeremy Straub, *Algorithmic Disgorgement Is Bad for Science and Society*, LAWFARE (June 12, 2023), https://www.lawfaremedia.org/article/algorithmic-disgorgement-is-bad-for-science-and-society [https://perma.cc/LLC7-73WX] (advocating for a regime in which algorithms and data are made available for public use rather than destroyed).

[224] When it comes to balancing the equities, a particular challenge arises in the context of machine learning models whose weights have been made publicly available (sometimes called "open source" or "open weights" models). *See* Off. of Tech., *On Open-Weights Foundation Models*, FTC (July 10, 2024),

C. *Alternative Remedies*

Most of the inquiries discussed so far involve comparisons. The proportionality inquiry compares the value of the model overall to the value attributable to the data underlying the defendant's violation. The relative hardship inquiry compares the burden that deletion poses to the defendant to the harm to the plaintiff posed by allowing the model to continue. But neither of these comparisons should occur in a vacuum—they must exist alongside considerations of other potential injunctive relief that the court could order. Courts issuing injunctive relief are directed "to grant relief no broader than necessary to cure the effects of the harm caused by the violation."[225] Courts considering model deletion should therefore weigh questions of proportionality and relative hardship alongside an assessment of whether there are alternative injunctions that would be less burdensome but adequately effective.

In many circumstances, the most relevant question will be whether there are feasible interventions to limit the use of the model in ways that would address the harms at issue. There are ways of controlling or influencing the use

https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models [https://perma.cc/Y89T-AWWW]. These models may be trained on large datasets by a sophisticated developer and then released for general use in ways that allow other individuals or enterprises to obtain and maintain their own copies of the model. *See, e.g.*, Mike Isaac, *How A.I. Made Mark Zuckerberg Popular Again in Silicon Valley*, N.Y. TIMES (May 29, 2024), https://www.nytimes.com/2024/05/29/technology/mark-zuckerberg-meta-ai.html [https://perma.cc/7NCE-4R2Z (staff-uploaded, dark archive)]. Because such models can become widely dispersed, it may be impossible for a court to effectively order their deletion. Court orders typically do not bind someone who was not party to the litigation, and so an order for Meta to delete its LLaMA model, for instance, would not reach the millions of copies of that model that are possessed by other individuals, or the tens of thousands of developers who have made their own software using the model. *See id.*; Taylor v. Sturgell, 553 U.S. 880, 892–95 (2008) (articulating the norm that non-parties are not bound by court judgments and noting limited exceptions). As a result, model deletion may not be an effective tool for remedying ongoing harm to a plaintiff in the context of an open-weights model.

It's not obvious what effect this would have on the balance of the equities in a case where the deletion of an open-weights model is sought. On the one hand, ongoing harm to a plaintiff might not be meaningfully resolved where it stems from a model with weights that have been widely disseminated—which could weigh against a model deletion order. On the other hand, it may be inequitable to allow a defendant to benefit from misconduct that they have made more severe by allowing the model to be disseminated. The answer is likely to be context-dependent, including an assessment of other options such as damages to compensate the plaintiff. And ultimately, addressing the harms of open-weights models may simply be beyond the capacity of individual lawsuits, with all of their limitations.

[225] City of N.Y. v. Mickalis Pawn Shop, LLC, 645 F.3d 114, 144 (2d Cir. 2011) (quoting Forschner Grp., Inc. v. Arrow Trading Co., 124 F.3d 402, 406 (2d Cir. 1997)); *see also* Rizzo v. Goode, 423 U.S. 362, 378 (1976) (noting "the settled rule that in federal equity cases the nature of the violation determines the scope of the remedy") (internal quotation marks and citation omitted).

of many types of models that can be implemented after the model is trained, and which may mitigate some harms.[226] This is especially true of the harms that may be associated with generative AI tools, such as the generation of copyrighted works or of private information. It may be possible to order a defendant to engage in an additional round of fine-tuning of its model that would influence its outputs but be less expensive than retraining the model in its entirety.[227] Or it may be possible to order the creation of a set of filters that prevents certain types of outputs from being generated, or that prevents the model from responding to certain types of inputs.[228] For many AI applications, the underlying trained model is a core part of the software tool, but it is "wrapped" in a user-friendly software system—and other features of that system, such as output filters, can be deployed to mitigate harms.[229]

Current work is also underway, exploring methods such as "machine unlearning" or "model editing" that may make it feasible to alter a model and address some kinds of harms without having to wholly retrain the model from scratch.[230] Machine unlearning, for instance, attempts to manage the problem of memorized information by revising the trained model to make it "forget" discrete information that it learned during the training process.[231] Model editing, in turn, attempts to edit more generalized knowledge that a model has learned during the training process, such as facets of an artist's style rather than memorizations of specific pieces of an artist's work.[232] These methods have the potential to save the considerable expense of retraining a model from scratch on

[226] *See* Paul Ohm, *Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI*, 25 COLUM. SCI. & TECH. L. REV. 214, 231–37 (2024) (describing interventions that may be done at different points in the production cycle of generative AI models) [hereinafter Ohm, *Fine-Tuning*].

[227] *See id.* at 223–28 (discussing fine tuning).

[228] *See id.* at 230–31 (discussing filtering).

[229] *See, e.g.*, Cooper & Grimmelmann, *supra* note 61, at 60 ("[S]ystem builders and operators have different places in which they can limit or prevent memorized content in models from being delivered to end users.").

[230] *See, e.g.*, Alessandro Achille, Michael Kearns, Carson Klingenberg & Stefano Soatto, *AI Model Disgorgement: Methods and Choices*, 121 PROC. NAT'L ACAD. SCI. 1, 3–4 (2024); Gandikota, *supra* note 61, at 3–4.

[231] *See* Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie & Nicolas Papernot, *Machine Unlearning*, IEEE (Dec. 15, 2020), https://ieeexplore.ieee.org/document/9519428 [https://perma.cc/4C7K-X8XC]; Aditya Golatkar, Alessandro Achille & Stefano Soatto, *Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks*, ARXIV 2 (Mar. 31, 2020), https://arxiv.org/pdf/1911.04933 [https://perma.cc/AV24-TQVM].

[232] *See* Gandikota et al., *supra* note 61, at 2 (distinguishing machine unlearning from model editing).

a new set of data that does not contain the infringing data in question.[233] But at least as of yet, the methods have significant flaws, not perfectly removing information and coming with significant tradeoffs to the model's performance after the intervention.[234]

A challenge of all of these approaches—unlearning, editing, filtering, and fine-tuning—is that, as Professor Paul Ohm has described, their efficacy may only be partial rather than absolute.[235] For generative AI models in particular, problems of explainability and interpretability often limit developers' ability to place reliable, guaranteed limits on their models' behavior.[236] Fine-tuning and filters can address many of the most common ways in which a given output is likely to be generated but leave open the possibility that the output will still be generated occasionally, particularly if there are determined users trying to elicit the output.[237] Even determined, good faith efforts may result in compliance that is probabilistic rather than perfect. And efforts to prevent certain outputs from being generated may have collateral consequences, imposing costs of money or time and impinging on the value of the model for other purposes.[238] Judges attempting to address the harms of generative AI tools via injunctions will have to manage difficult tradeoffs of cost and accuracy and may have to develop standards that accommodate imperfect compliance and learning over time.[239]

Another potential avenue of approach in some cases would be to impose a kind of mandatory license. Where the underlying harm is an intellectual property violation, for instance, courts may be able to require a defendant to pay a fee to the plaintiff each time the model generates an output that bears sufficient resemblance to the IP in question. This is easier said than done. Doing so would require the court to have confidence that it is possible to reliably determine how often the relevant material is generated, which may be difficult.

[233] *See* Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Am. Smith & Chiyuan Zhang, *MUSE: Machine Unlearning Six-Way Evaluation for Language Models*, ARXIV 2 (July 14, 2024), https://www.arxiv.org/pdf/2407.06460 [https://perma.cc/JU46-9UY7].

[234] *Id.* at 2–3 (finding that "unlearning algorithms generally fail to meet data owner expectations in preventing privacy leakage," and describing ways that applying unlearning algorithms can degrade model performance).

[235] Ohm, *Fine-Tuning*, *supra* note 226, at 238–39.

[236] *See id.*; *see also* Jessica Ji, Josh A. Goldstein & Andrew J. Lohn, *Controlling Large Language Model Outputs: A Primer*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY, at 5–10 (2023) (describing methods for controlling language model output and their limitations).

[237] Ohm, *Fine-Tuning*, *supra* note 226, at 238–39.

[238] *Id.*

[239] *See id.*

If, for instance, a model generates text that is ninety percent identical to a copyrighted text, is that enough to trigger the license? Eighty percent? Does the threshold depend on which twenty percent is different, or how it is different?[240] Purely numeric thresholds may be easier to administer, but it may be difficult to establish a threshold that is not arbitrary. Still, mandatory licenses have worked in other contexts, and despite their flaws they may still be preferable to ordering the deletion of a model in drastically disproportionate scenarios.[241]

One benefit of a mandatory licensing approach is that it may be able to compensate for some of the shortcomings of injunctions that directly try to block the generation of the unwanted outputs via fine tuning or filters. Rather than attempting to directly control a defendant's use of technology to achieve a desired end, a licensing regime creates a financial incentive for the defendant to reduce the number of times an output is generated without specifying the technical means. The efficacy of such a regime would depend not only on detection but also on setting the right level of fee; courts may wish to consider setting an explicitly supercompensatory fee if the goal is deterring the generation of the undesired output rather than simply compensating the plaintiff for it.

For all of these approaches, an additional and central challenge is that they are likely to require some amount of technical expertise in a domain where technology is rapidly changing. Courts may be understandably wary of their ability to superintend a company's technical safeguards on its model's outputs, including their ability to referee a battle between expert witnesses presented by the plaintiffs and the defendants. Approaches like the license-as-incentive regime described in the previous paragraph can reduce the burden of technological expertise somewhat but still requires enough know-how to establish a system that determines what outputs are relevant and detects them reliably enough to trigger the licensing fee.

The costs and difficulty of administering a remedial regime are a legitimate consideration when deciding which remedy to impose. As a result, there might be some situations where model deletion is the preferable remedy

[240] For instance, the specific ways in which an image has been altered may be relevant to a fair use defense, such as if the altered image qualifies as a parody. *See, e.g.*, Sag, *supra* note 66, at 1901–02.

[241] A more global solution to intellectual property issues with generative AI tools may require legislation. *See, e.g.*, Frank Pasquale & Haochen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis*, 110 U. VA. L. REV. ONLINE 207, 208–211 (2024). If such a legislative regime is created, that would likely have significant implications for model deletion and any of the alternative remedies discussed here.

simply because it is easier to administer than other alternatives. But these alternatives may become easier to administer over time, as experience with this technology becomes deeper and more widespread, and as courts see more of these cases. And as the value of machine learning tools continues to grow, it will be more necessary for courts to consider alternatives to deletion if they desire to implement fair and proportionate remedies.

CONCLUSION

Model deletion has emerged as a powerful new remedy in the world of data and AI regulation. As this Article has shown, it offers distinct advantages in addressing ongoing harms and providing effective deterrence against unlawful data practices. But it also has some significant potential flaws. In its current formulation, model deletion threatens to be a disproportionate penalty in a variety of easy-to-imagine circumstances.

It should be possible to strike a balance between model deletion's benefits and limitations. As the remedy is sought more often, courts and public enforcers should look to traditional equitable factors to guide their assessments of when to use model deletion, conscientious of its strengths and weaknesses. The answer in a given case will not always be easy or obvious—just like many other legal doctrines. But as the regulation of AI tools becomes more and more important for commerce and governance, it will likewise be increasingly important to harness the benefits of model deletion while mitigating its downsides.