Ordering Model Deletion

In litigation over both predictive and generative AI tools, a new remedy has earned a significant amount of attention and praise. Sometimes called "algorithmic disgorgement" or "algorithmic destruction," this remedy entails deleting not only the data used to train a machine learning model, but also deleting the model itself. This remedy of model deletion has earned praise from enforcers and commentators, has been increasingly used in enforcement actions by the FTC, and is also now being sought in a variety of private lawsuits such as the New York Times' suit against Microsoft and OpenAI.

But the relatively narrow contexts in which model deletion has been used so far mask a serious shortcoming: if algorithmic disgorgement is extended to new contexts such as the New York Times suit, it will likely be a grossly disproportionate remedy. Such an extension is not supported by the law of disgorgement, nor by the traditional equitable test governing permanent injunctions (which would presumably apply to an order for model deletion, and which is incorporated in certain intellectual property contexts as well). This paper, which is an early-stage law review submission, will have two goals: first, to articulate this flaw in model deletion as a remedy, and second, to help bring some "order" to model deletion orders by exploring how to determine where the remedy will and will not be appropriate.

First, model deletion's potential for disproportionality represents a serious shortcoming that will likely arise in many contexts. As the FTC has used the remedy so far, it has applied to models derived "in whole or in part" from data that has been unlawfully obtained or retained by a defendant company. Such an approach makes some sense in the cases in which the FTC has used it, such as the Cambridge Analytica case or Everalbum case, where the machine learning models involved were relatively specialized tools in which the unlawful data at issue formed a key part of the tools' creation. In these cases, the models that were ordered to be deleted appear to have derived their value in large part from the unlawful conduct that formed the basis for the FTC's action.

But that will not always be the case where model deletion is sought. 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. Some machine learning tools, such as large language models in particular, train on enormous quantities of data and have wide ranges of commercial uses. In such scenarios, small quantities of data may still be responsible for some of the value embodied in the model at issue, but they often will not account for a significant fraction of that model's total value. Where individual litigants seek to vindicate their rights with respect to a small quantity of the overall training data, model deletion will be a disproportionate remedy: it will ask for the destruction of an asset whose value is far beyond the harm to the plaintiff.

Where model deletion is significantly disproportionate, it is unsupported by existing theories of disgorgement and unjust enrichment, and likely would not meet the broad equitable standard for a permanent injunction. The goal of disgorgement as a remedy is generally "to eliminate profit from wrongdoing while avoiding, so far as

possible, the imposition of a penalty."[1] To accomplish this goal, the unjust enrichment of a wrongdoer is generally set to be the higher of either "the reasonable cost of a license" to the asset that was wrongfully appropriated or "the net profit attributable to the underlying wrong."[2] Although the attribution of value in the context of large language models will be difficult (and perhaps impossible to achieve with a high degree of precision), it is unlikely that a rigorous account would, for instance, put OpenAI's net profit attributable to training on the work of the New York Times as equivalent to the existence of GPT-4 as a whole. As a result, an order of model deletion may be appropriate in group litigation or in an action by a public enforcer where the unlawful data collectively accounts for a significant portion of the total value of a trained model, but in other scenarios it will often be inappropriate.

Given that there are relatively clear scenarios where model deletion will be appropriate and also ones where it will not be appropriate, how should courts and other enforcers go about deciding what to do? Existing law provides little to no framework other than the vague and general contours of the test for injunctive relief. This paper will attempt to color in some of the details of that test, and will argue that a decision-maker such as a court should examine the following factors: First, the court should evaluate the nature of the harms alleged, with particular attention to whether the harms are ongoing, their likely scale and frequency, and whether they are the kinds of harms that are traditionally well compensated via money damages. Second, the court should assess the value of the trained model at issue as well as the cost of replacing that model with a new one that is not trained on the underlying unlawfully possessed data. Third, the court should then attempt at least a rough approximation of how much of the value in the model is attributable to its training on the unlawful data at issue.

In scenarios where there is ongoing harm and the value of the model is largely accounted for by the unlawful data, model deletion is a reasonable approach. If not, the court should make two more inquiries: whether there are alternative measures, such as filtering, fine-tuning, or machine unlearning, that could address a substantial fraction of the ongoing harm; and whether there are other circumstances that would justify a disproportionate remedy, such as a defendant's willfulness or a likely predominance of unlawful material in the training data for the model beyond the data at issue in the case at hand. After considering these factors, the court may determine that model deletion is justified, or may hold that model deletion would be unjustifiably disproportionate relative to the defendant's wrongdoing and resultant gain.

By sketching out these proposed contours, I hope this paper will contribute to the ongoing conversation about a nascent but important remedy in the AI space. And in doing so, I aim to build on the existing literature around remedies in the AI context to emphasize that fair and effective AI regulation will depend not only on getting substantive liability standards right, but also on designing procedures and remedies that attend to the nuances of this area of law and technology.

---

[1] Restatement (Third) of Restitution and Unjust Enrichment § 51 (2011).
[2] *Id.*

Selected Bibliography

Lucas Bourtoule et al, *Machine Unlearning*, IEEE Symposium on Security and Privacy (2021), https://arxiv.org/pdf/1912.03817

Ryan Calo, *Privacy Harm Exceptionalism*, 12 Colo. Tech. L.J. 361 (2014)

Bernard Chao, *Privacy Losses As Wrongful Gains*, 106 Iowa L. Rev. 555 (2021)

Emma Elder, *Wrongful Improvers As A Guiding Principle for Application of the Ftc's Ip Deletion Requirement*, 97 Wash. L. Rev. 1009 (2022)

Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models As the Ftc's Newest Enforcement Tool*, 29 Rich. J.L. & Tech. 1 (2023)

Jevan Hutson & Ben Winters, *America's Next "Stop Model!": Model Deletion*, 8 Geo. L. Tech. Rev. 124 (2024)

Ruoxi Jia et al, *Towards Efficient Data Valuation Based on the Shapley Value*, Proc. 22nd Int'l Conf. on A.I. and Statistics (2019), https://proceedings.mlr.press/v89/jia19a.html

Katherine Lee, A. Feder Cooper & James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, J. Copyright Society of the U.S.A. (forthcoming 2024)

Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. Rev. 479 (2022)

Paul Ohm, *Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI*, Colum. Sci. and Tech L. Rev. (forthcoming)

Pamela Samuelson et. al., *Recalibrating the Disgorgement Remedy in Intellectual Property Cases*, 100 B.U. L. Rev. 1999 (2020)

Lauren Henry Scholz, *Privacy Remedies*, 94 Ind. L.J. 653 (2019)

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 (2021)

Jinsung Yoon, Sercan O. Arik, and Tomas Pfister, *Data Valuation using Reinforcement Learning*, Proc. 37th Int'l. Conf. on Machine Learning (2020), https://proceedings.mlr.press/v119/yoon20a.html

Eric Zhang et al, *Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models*, arXiv (Mar. 2023), https://arxiv.org/abs/2303.17591

Boxin Zhao et al, *Addressing Budget Allocation and Revenue Allocation in Data Market Environments Using an Adaptive Sampling Algorithm*, arXiv (June 5, 2023)