From *Algorithmic Destruction* to *Algorithmic Imprint*: Generative AI and Privacy Risks Linked to Potential Traces of Personal Data in Trained Models

Lydia Belkadi [*] [1] **Catherine Jasserand** [*] [1]

Abstract

This contribution discusses the 'algorithmic disgorgement' tool used by the FTC in four settlement cases relating to unfair or deceptive data practices, where the FTC ordered to delete not only the data unlawfully processed but also the resulting models. According to some scholars, this measure could mean and be justified by the fact that some models contain traces or shadows of training data. Reflecting on this tool from a USA and EU legal perspective, we question the opportunity to design more granular legal assessments and contend that regulators and scholars should consider, if evidenced, whether (traces or fragments of) personal information can be contained in or disclosed by models when defining deletion or retraining obligations. This issue has received limited interdisciplinary attention, hindering ongoing discussions on generative AI regulation.

For a few years, the US Federal Trade Commission ('FTC') has developed a new enforcement tool in several privacy settlement cases where personal data were illegally and wrongfully collected to train models or algorithms. In these settlements, the FTC ordered the companies at stake to not only delete the data collected for training purposes but also the resulting algorithms. Known as 'algorithmic disgorgement', this mechanism has been used against several companies, including the photo-sharing app Everalbum (FTC, 2021). The FTC has recently received a complaint to examine Open AI's practices, and particularly one of its foundational AI models, the Generative Pre-Transformer 4 ('GPT-4'). The complaint emphasized that the company's Usage Policy is constantly changing and waives liability for unlawful, deceptive, unfair, and dangerous applications. It also points out that Open AI did not implement satisfactory deterrence measures to protect consumers, merely relying on a disclaimer on possible misuses (CAIDP, 2023). The FTC has not yet decided whether it will follow-up on the complaint.

1. What is 'Algorithmic Disgorgement'?

The Federal Trade Commission ('FTC') is the US regulator tasked with protecting consumer rights and preventing market distortions. Following Section 5 of the FTC Act, the FTC has a competence to regulate 'unfair or deceptive acts or practices', including privacy harms. While unfairness constitutes a substantial consumer injury that is not reasonably avoidable and not outweighed by countervailing benefits for the consumer, deception is characterized as a representation, omission or practice that is likely to mislead a consumer where that representation or interpretation is material and based on a reasonable consumer standard (Hoofnagle, 2016). The FTC has limited authority to impose monetary damages and must exercise its powers under narrowly defined and constrained procedures (Hoofnagle, 2016). A recent decision of the Supreme Court reiterated these limitations and condemned the FTC's use of injunctive relief to obtain monetary penalties (Hoofnagle, 2016; Goland, 2023).

The FTC developed a new enforcement tool known as 'algorithmic disgorgement' in several privacy settlement cases where personal data were illegally and wrongfully collected to train models or algorithms. Algorithmic disgorgement is a non-monetary enforcement mechanism aiming to renew and augment FTC's ability to regulate algorithmic consumer harms (Slaughter, 2021).

In the four settlements in which the FTC applied this mechanism, the authority found the existence of unfair or deceptive practices harming consumers in different contexts. In the settlement with Cambridge Analytica, the company had harvested personal data from Facebook users by misrepresenting which data was collected and for which purpose (FTC, 2019). In Everalbum, the FTC found that the company's conditions for using facial recognition technology and exercising control over the technology were unfair or deceptive (FTC, 2021). In Weight Watchers, the company had harvested children's personal and sensitive health data without

[*] Equal contribution [1] Center for IT & IP Law, KU Leuven University, Leuven, Belgium. Correspondence to: Lydia Belkadi <lydia.belkadi@kuleuven.be>, Catherine Jasserand <catherine.jasserand@kuleuven.be>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

prior notice and parental consent (FTC, 2022). Finally, in Ring, the company had deceived its customers by allowing employees and contractors access to the customers' private videos and using the videos to train algorithms without the customers' consent (FTC, 2023).

In these four settlements, the companies were ordered to delete not only the data used for the training but also the resulting algorithms (described as 'affected work product' in the settlements). The mechanism used by the FTC is conceived as 'a penalty the agency can wield against companies that used deceptive data practices to build algorithmic systems' (Protocol, 2022). If the FTC considers that the illegality of the data collection and processing contaminates the product obtained (i.e. the trained models), there is no indication that algorithmic disgorgement's rationale is based on traces of personal data that might be contained therein. Instead, this mechanism seems to be similar to the 'fruit of the poisonous tree' doctrine (Federman, 2021).

2. An Enforcement Mechanism Linked to Data Shadows in Trained Models?

The FTC's settlements have not been challenged in court, thus there are limited information available regarding the legal reasoning used to design this enforcement mechanism. In turn, many scholars have examined different legal bases and interpretations. Focusing on model deletion, US scholars have explored the possibility that the FTC applied algorithmic disgorgement due to the traces of personal data retained by models (Li, 2022). This thesis has generated important discussions amongst US scholars on the relationship between training data and trained models. Some authors went a step further, claiming that removing or deleting an algorithm is insufficient as the harm caused might survive its destruction (Ehsan et al., 2022). A similar line of arguments was developed in the CAIDP complaint, considering that Open AI's generative model was wrongfully trained on and contains personal data (CAIDP, 2023).

Discussions are increasingly being held in Europe about the impact of withdrawing specific data from a dataset and whether a model should be retrained. In the context of the European data protection regime (i.e., 'General Data Protection Regulation' or 'GDPR'), legal scholars have examined for a few years the scope of selected obligations and individual rights, as well as their implications for models. For example, some authors have raised interrogations on the information retained and potentially revealed by models while applying and assessing the principles of data integrity and confidentiality, as well as the obligation of security. Focusing on different attacks that may lead to privacy breaches, these authors discussed whether models could be considered personal data (Veale et al., 2018). Others have examined the applicability of the principle of accuracy and the requirement of data quality to processing operations (Hallinan & Borgesius, 2020; Dimitrova, 2021), and argued that the right to rectification could extend to the underlying data processing technology (Dimitrova, 2021).

3. Does a Model Contain Traces or Fragments of Personal Data?

These legal discussions in the US and the EU have taken place in parallel to and in support of essential strands of technical research that question Machine Learning and Deep Learning models' privacy risks linked to unintended memorization, training data leakages, membership inference or model inversion attacks, and the potential to design 'unlearning' mechanisms'.

However, the legal continuum between training data and models formulated by the FTC may prove counterproductive if not associated with further contextual assessment. For example, it has been argued that FTC's orders adopted a generic terminology and remedy, while each case followed its own 'fact pattern' (Goland, 2023). These orders relate to the 'affected work product' defined broadly as any models or algorithms developed in whole or in part using data acquired through unfair and deceptive data collection practices. However, as signaled elsewhere, this formulation results in a 'full-scale destruction of all models obtained from tainted data', which might restrain the development and implementation of deletion measures (Goland, 2023).

Furthermore, legal scholars have extended the analysis of algorithmic disgorgement beyond the effective scope of the FTC's enforcement mechanism. The hypothesis that this enforcement mechanism applies to 'data shadows' calls for two remarks. First, the FTC did not state that they ordered the deletion of the models due to the potential traces of personal data that the models contained. One could indeed argue that the algorithmic disgorgement results from a combination of competition (deceptive practices) and privacy violation (unlawful processing of personal data). The deletion of models is a logical result, based on deceptive practices, independently of whether a model can contain traces (or shadows) of personal data. Second, the legal scholars who have interpreted algorithmic disgorgement have not demonstrated or relied on any research or experiments that could disclose the content of a model. Yet, to draw conclusions on whether a model should be deleted or retrained when personal data are withdrawn from a dataset, one should determine the effects of such withdrawal on the content of models.

Thus, legal communities are increasingly relying on relatively generic, open or vague legal terminologies, while such technical and legal assessments require a precise understanding of design choices and their trade-offs. In turn, this situations suggests a need to further anchor enforcement mechanisms to scientific evidences, as well as tangible mitigation and corrective measures tailored to specific contexts. Identifying what information is contained and revealed by a model can only be established through an interdisciplinary process that will help assess whether the enforcement mechanisms deployed in the USA and in Europe are adequate.

4. Is Data Deletion or Model Retraining an Adequate Solution?

The effects of data deletion on downstream tasks remain insufficiently understood by legal communities. Legal analyses remain confused regarding the nature and means for identifying the link, if any, between the training data and trained models. Yet, ascertaining the existence of such link is a necessary preliminary step to reflect upon new regulatory interventions, secure a reasonable degree of legal certainty and explore potential corrective measures throughout models' life cycles (e.g., ex ante and ex post assessments).

This interdisciplinary gap hinders discussions on more granular and contextual approaches to legal assessments – not all AI models are born equal. Identifying risks and mitigation measures are two crucial elements to guide the legal assessment of whether, and under which conditions, regulations should impose an obligation to retrain models without the data at stake. In this context, very specific and practical legal questions may be raised – e.g., Should data deletion extend to the retraining of models under specific conditions (such as the type and quality of personal data at stake) and in compliance with the principle of proportionality (meaning that not every data deletion should lead to retraining the model)? What are the privacy risks associated with disclosure of data? Should a model be considered data accurate if it still contains traces of deleted data? What is the relationship between training data and artificial hallucinations? Could a generative model leak data that have been deleted from the training datasets, and if so, under which conditions?

5. Conclusion

Many challenges remain regarding the legal grounds to impose obligations to delete or retrain models. This points to the importance of determining more precisely whether and to what extent models contain (fragments or traces of) personal data. This assessment is not only necessary to assess privacy and security risks, but also to evaluate the enforcement mechanisms and potential new technical-legal venues to protect individuals as data subjects and consumers. Further interdisciplinary research will be necessary on this important emerging topic.

Acknowledgements

The authors express their gratitude for the insightful comments provided by the anonymous reviewers.

Lydia Belkadi, doctoral researcher, funded by European Commission's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 (TReSPAsS).

Catherine Jasserand, postdoctoral researcher (IF-MSCA), funded by the European Commission's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No. 895978 (DATAFACE).

References

Federal Trade Commission Act, Section 5. *15 U.S.C. §§41-59.*

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ec (General Data Protection Regulation). *OJ L 119, 4.5.2016, 1-88, data.europa.eu/eli/reg/2016/679/oj/.*

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. *COM/2021/206 final, www.eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:520221PC0206/*, 2021.

CAIDP. In the matter of OpenAI (FTC 2023). *Center for AI and Digital Policy (30 March 2023), www.caipd.org/cases/openai/*, 2023.

Dimitrova, D. The Rise of Personal Data Quality Principle. Is it Legal and Does it Have an Impact on the Right to Rectification? *European Journal of Law and Technology*, 12(3):1–31, 2021.

Ehsan, U., Singh, R., Metcalf, J., and Riedl, M. O. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability and Transparency*, pp. 1305–1317, New York, NY, USA, 2022. Association for Computing Machinery.

Federman, H. Tainted fruit: Disgorgement of data from the FTC and beyond. *iapp (27 April 2021), www.iapp.org/news/a/tainted-fruit-disgorgement-of-data-from-the-ftc-and-beyond/*, 2021.

FTC. Cambridge Analytica, LLC, *In re. Common File No. 182 3107 (24 July 2019, 6 December*

2019) www.ftc.gov/legal-library/browse/cases-proceedings/182-3107-cambridge-analytica-llc-matter/, 2019.

FTC. Everalbum, Inc., *In re*. *Common File No. 1923172 (11 January 2021, 7 May 2021) www.ftc.gov/legal-library/browse/cases-proceedings/192-3172-everalbum-inc-matter/*, 2021.

FTC. Weight Watchers/WW. *Common File No. 192 3228 (16 February 2022, 4 March 2022) www.ftc.gov/legal-library/browse/cases-proceedings/192-3228-weight-watchersww/*, 2022.

FTC. Ring LLC. *Common File No. 202 3113 (31 May 2023) www.ftc.gov/legal-library/browse/cases-proceedings/2023113-ring-llc/*, 2023.

Goland, J. A. Algorithmic Disengorgment: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data. *Richmond Journal of Law and Technology*, 29(2):1–51, 2023.

Hallinan, D. and Borgesius, F. Z. Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle. *International Data Privacy Law*, 10(1):1–10, 2020.

Hoofnagle, C. J. *Federal Trade Commission Privacy Law and Policy*. Cambridge University Press, 1st edition, 2016.

Kaye, K. The FTC's new enforcement weapon spells death for algorithms. *Protocol (14 March 2022), www.protocol.com/policy/ftc-algorithm-destroy-data-privacy/*, 2022.

Kranenborg, H. Article 17 Right to erasure ('right to be forgotten'). In Kuner, C., Bygrave, L., Docksey, C., and Drechsler, L. (eds.), *The EU General Data Protection (GDPR): A Commentary*, chapter 3, pp. 475–484. OUP, 2020.

Li, T. Algorithmic Destruction. *SMU Law Review*, 75(3): 479–506, 2022.

Protocol. How to kill an algorithm. *Protocol (17 March 2022), www.protocol.com/newsletter/protocol-enterprise/ftc-algorithmic-disgorgement-japan-chips/*, 2022.

Slaughter, R. K. Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission. *Yale Journal of Law and Technology*, 23:1–63, 2021.

Veale, M., Binns, R., and Edwards, L. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, 376:1–15, 2018.