Model Destruction – The FTC's Powerful New AI and Privacy Enforcement Tool

March 22, 2022

A recent FTC settlement is the latest example of a regulator imposing very significant costs on a company for artificial intelligence ("AI") or privacy violations by requiring them to destroy algorithms or models. As companies invest millions of dollars in big data and AI projects, and regulators become increasingly concerned about the risks associated with automated decision-making (e.g., privacy, bias, transparency, explainability, etc.), it is important for companies to carefully consider the regulatory risks that are associated with certain data practices. In this Debevoise Data Blog post, we discuss the circumstances in which regulators may require "algorithmic disgorgement" and some best practices for avoiding that outcome.

How the Model Destruction Remedy Arises

This issue can arise when a regulator is scrutinizing a complex model that has been trained or enhanced using data that the owner was not legally authorized to use for that purpose. Examples could include:

- Company A built a model to screen resumes and decide which job applicant gets to the interview stage of the process. The model was trained using the resumes of current and former employees without their knowledge or consent;
- Company B built a model to review loan applications and decide who is an unacceptable credit risk, based in part on data that was scraped from the Internet in violation of the terms of use of certain websites; and
- Company C built an AI model that can identify false news stories. The model was trained using verified news articles from reputable sources, but Company A did not obtain the necessary licenses and copyrights to use the news articles for that purpose.

If it is determined that an AI model was developed using data that was not legally permitted for that purpose, two questions arise. First, can the tainted data be removed from the model entirely, or does the nature of the model preclude that possibility? And

second, even if the model can be completely cleansed of the tainted data, should the appropriate remedy for the data violation include destruction of the model?

W.W. International & Kurbo

On March 3, 2022, the FTC reached a court-approved settlement with Kurbo, Inc. and W.W. International, formally Weight Watchers International (collectively, "Weight Watchers"). The FTC's complaint alleged that Weight Watchers had collected and retained children's personal information without the necessary notices and consents in violation of the Children Online Privacy Protection Act ("COPPA") and Section 5 of the Federal Trade Commission Act. The settlement included injunctive relief requiring Weight Watchers to delete or destroy any personal data collected online from children under the age of 13 without verifiable parental consent, as well as any models or algorithms that were developed using that personal information. The settlement also prohibits Weight Watchers from "disclosing, using or benefitting from" any of the personal information obtained prior to the settlement date, absent verifiable parental consent. As for monetary relief, Weight Watchers also agreed to pay a $1.5 million civil monetary penalty for various data violations including "retaining personal information collected online from a child for longer than reasonably necessary to fulfill the purpose for which the information was collected." Although the FTC is not authorized to obtain civil penalties for initial violations of the FTCA, the FTC can obtain civil penalties for initial violations of other statutes it enforces, such as COPPA.

Everalbum

We first wrote about model destruction in early 2021, in connection with the FTC's requirement that Everalbum delete its facial recognition algorithms that were developed using the photos and videos of customers without their consent. The company had allegedly violated Section 5(a) of the FTC act by promising customers that it would only use facial recognition on users' content if they opted in, and that it would delete users' content if they deactivated their account but did not adhere to either of those promises. In announcing the settlement, Former FTC Commissioner (and current Director of the Consumer Financial Protection Bureau), Rohit Copra, stated: "First, the FTC's proposed order requires Everalbum to forfeit the fruits of its deception. Specifically, the company must delete the facial recognition technologies enhanced by any improperly obtained photos. Commissioners have previously voted to allow data protection law violators to retain algorithms and technologies that derive much of their value from ill-gotten data. This is an important course correction."

The Everalbum case was not the first time that the FTC had required a company to delete algorithms it had created with data it did not have permission to use in that way. In 2019, the FTC ordered Cambridge Analytica to destroy algorithms derived from information collected from consumers without the necessary notices and consents.

Why Algorithmic Disgorgement Is a Potent Remedy

For the FTC in particular, model destruction is a powerful new tool, particularly with respect to alleged violations involving AI technologies. Due to the recent Supreme Court decision in AMG Capital Management, the FTC is precluded from obtaining equitable monetary remedies pursuant to Section 13(b) of the Federal Trade Commission Act (FTCA). The FTC, however, can still obtain a wide array of non-monetary equitable remedies, which can, in certain circumstances, be even broader in scope than the conduct deemed unlawful. These are known as "fencing-in" remedies and are intended to prevent future unlawful conduct.

In addition, the vast majority of FTC cases result in settlement, and the FTC can often obtain remedies via settlement (including stipulated injunctive remedies) that it might not be able to obtain based on a court order alone. In light of the FTC's recent settlements, it is clear that the FTC intends to pursue algorithmic disgorgement remedies unless and until a court rules that such a remedy exceeds the Commission's authority.

For all regulators, model destruction may serve as a strong punishment and a powerful deterrent because:

- Many complex models have taken years to develop, at the cost of millions of dollars, and cannot be easily replicated or replaced;
- Rich data sets are often used to train multiple models, so if the data sets are tainted, several models may be impacted; and
- Even if tainted data is only used to train one model, the outputs of many models serve as the inputs for other models, so algorithmic disgorgement may, in some cases, require the destruction of several models in the same model chain or cluster.

As a result, companies should consider taking steps to reduce the risk of engaging in violations that could lead to the imposition of this powerful remedy.

Takeaways and Tips for Avoiding Model Destruction

- Don't Wait for AI Regulation to Build Model Governance and Compliance
    - As we have noted previously, the SEC, the FTC and other regulators are not waiting for new AI-specific regulations to bring enforcement actions related to the use of complex models. The FTC alleged very traditional violations in these cases—that the companies engaged in misrepresentations regarding the use of customer data for training or operating models, and such misrepresentations

constitute unfair or deceptive acts or practices in violation of Section 5(a) of the FTC Act, which was passed in 1917.

- Identify High Risk Models
  - Consider creating a system for identifying significant models or clusters of interconnected models that may contain inputs that would put the models at risk for destruction and have a means by which these models can be reviewed for compliance and risk.
- Review High-Risk Models for Rights to Use Inputs for the Model
  - Consider implementing policies, procedures and training designed to ensure that the company is (1) aware of the inputs that are being used for the training and operation of its high-risk models and (2) authorized to use the data for those purposes. Consider conducting sample audits to make sure that any necessary notices are provided and appropriate consents are obtained.
- Business Continuity Planning
  - Consider creating a plan to ensure continued operations, without significant interruption, if particular high-risk models were ordered to be destroyed or failed for some other reason. The short time periods that the FTC provided for the destruction (90 days or less) suggest that companies may need to move quickly to replace models that regulators determine are ill-gotten gains from unlawful use of data.
- Track What Data Is Used to Train Significant Models
  - To the extent that a certain set of training data is determined to be tainted by a regulator, it may become important for the company to prove that other models were not trained using that same data set. Accordingly, companies should consider having robust documentation for the kinds of data that was used to train, validate and operate significant models.
- The Need for Vendor and Acquisition Diligence
  - Companies are increasingly bolstering their AI capabilities through acquisitions or third-party vendor arrangements. In light of the risk of data violations, companies should consider implementing a robust AI diligence and risk-assessment process for significant AI applications or data sets that were developed by third parties, or are to be acquired, that could include:

  - Determining whether the AI application was developed using sensitive consumer data—including biometric information or data concerning protected class membership—or other data that may be subject to claims of unauthorized use;

  - Assessing what steps the vendor or acquisition target took to ensure that all the appropriate notices were provided, and authorizations were obtained; and

  - Evaluating the documentation associated with those authorizations.

- Consider Ways to Mitigate Risks and Costs Related to Tainted Models

  - To the extent that a company has already collected and used data for model training that may be viewed as problematic, efforts should be made to determine whether any remediation is possible by:

    - Providing appropriate after-the-fact notices of the data's use or obtaining necessary consents;

    - Completely purging the potentially tainted data from the model, to the extent possible, and documenting that process, perhaps with the assistance of a third-party firm to provide some testing or audit capabilities;

    - Planning for any business disruption that would result if the company were obligated to temporarily or permanently cease use of the model;

    - Ensuring that the risk that the model may have to be scrapped due to tainted training data is sufficiently disclosed to the board and investors; and

    - Assessing whether the risks can be further mitigated with insurance or otherwise.

Conclusion

These FTC settlements, and the prospect of algorithmic disgorgement, have significant implications for companies that rely on consumer data to train and operate significant AI applications or that license AI from third parties.

Regulators are focused on AI fairness, bias and privacy concerns and are increasingly likely to find instances of data being used for model development in a way that the regulators find objectionable. When that happens, extracting that noncompliant data from algorithms may be extremely difficult, especially for machine learning or AI

models, rendering destruction a logical remedy for regulators. Accordingly, companies that are heavily investing in AI and big data should consider implementing policies, procedures, training and governance to reduce that risk.

The authors would like to thank Debevoise law clerk Emily Harris for her contribution to this post.

To subscribe to our Data Blog, please click here.

* * *

Please do not hesitate to contact us with any questions.

NEW YORK

Avi Gesser
agesser@debevoise.com

WASHINGTON, D.C.

Paul D. Rubin
pdrubin@debevoise.com

NEW YORK

Anna R. Gressel
argressel@debevoise.com