AI Unlearning as a Powerful Legal and Business Tool for the AI Industry

Improper data collection and use to train AI engines can raise significant legal and ethical risks, whether by violating a website's T&Cs through data scraping, by breaching copyright laws or confidentiality agreements, through non-compliant handling of personal data, or through compromised AI applications that breach applicable laws or ethical standards. The issue of compromised input datasets is particularly challenging for the AI industry because neural networks evolve structurally with the compromised data, and reversing this learning is challenging: the effects of the compromised training data may remain a permanent part of the AI model even as it evolves with additional, uncompromised data. These issues can be further amplified by second-order AI models that rely on compromised first-order models, raising complex legal queries on liability. Not only could the AI vendors be exposed to liability, but legal and business risks could also propagate to the customers and users of such compromised AI engines.

Consequently, the ability for an AI engine to "unlearn" compromised data could be a powerful solution for the AI industry and could resolve multiple legal and ethical issues across the whole AI vendor and user chains.

1. Data used to train an AI engine can raise material legal issues

There are numerous legal issues that can arise when improper data is used to train an AI engine, including the following:

(a) Content could be scraped from websites in violation of website T&Cs. Most websites prohibit data scraping, so this is a universal concern for data scraping.

(b) Content could be collected and used without permission of content owner. This could be a violation of copyright protection (e.g., copyright law automatically provides to the content owner the exclusive rights to reproduce and produce derivative works of content). Defenses along the line of fair use are possible to assert, but become more complicated as the scope of data used to train the AI model increases.

(c) Content could be used in violation of confidentiality agreements in effect between the AI engine vendor and the owners of the content (e.g., a cloud services provider that uses a customer's data stored in the cloud platform to train an AI engine in violation of confidentiality restrictions in the applicable SAAS agreement with the customer).

(d) Personal data could be used to train AI engines without obtaining the appropriate consents from individuals. This could trigger automatic breaches under privacy laws and under other laws and regulations (e.g., Credit Card Network Rules, HIPAA, other US Federal laws).

(e) In some cases, generative AI models were trained using the output from AI engines that themselves were trained using compromised training data sets (i.e., second order AI engines trained using outputs from first-order compromised AI engines). This triggers some interesting legal questions, including whether the second-order AI engines are themselves compromised, and whether end users of the second-order AI engines could be exposed to legal liability. We will address these topics in the future.

(f) Data could be selected in a way that triggers impermissible biases in AI engine actions (e.g., collecting and using data in a way that causes the AI engine to make discriminatory financial decisions in violation of financial and consumer protection laws).

2. Unique Data Issues for the AI Industry

The fundamental problem with using compromised data to train an AI engine is that the neural network itself is modified by the data (e.g., the weights evolve based on the training data), so there is no obvious way to reverse the learning. For example, as more and more uncompromised data is used to continue to train the engine and the neural network evolves further, the impact of the compromised data is attenuated, but is not completely removed.

We already see a number of prominent law suits launched against AI engine vendors, where content owners are alleging that their content was used to train AI engines without their consent. These disputes will undoubtedly increase in number and scope, and may also propagate across the AI vendor and user industry segments.

3. Unlearning Could be a Powerful Solution for the AI Industry to Address Compromised Data

Considering the risks outlined above, the ability to undo learning from compromised data could be a powerful solution to mitigate the consequences of compromised training data sets. If an AI engine could be modified to remove the effect of the compromised data, the legal and business ramifications could be momentous for both AI vendors and AI customers.

The following are some areas where AI "unlearning" could have material impacts:

(a) The legal breach that occurred when the compromised data was initially collected and used without permission would still remain (i.e., cannot undo past copying of content without the permission of the content owner), and the content owner could still be entitled to some damages (e.g., statutory damages for copyright infringement). But the risk that damages would be awarded to the content owner based on the profits of the vendor of the compromised AI engine could be largely avoided if the neural network can be modified to reverse the modifications induced by the compromised data. Alternatively stated, if the impact of the compromised data on the AI engine is removed, none of the subsequent profits of the AI engine vendor could be attributed to the infringing activity any longer.

(b) A copyright owner could no longer seek a preliminary or permanent injunction against an AI engine vendor because there would be no future or ongoing copyright infringement once the compromised data is unlearned.

(c) In general, in the absence of a way to unlearn compromised content, once a content owner can prove that compromised data was used to train an AI engine in violation of the owner's rights, the content owner acquires significant leverage in litigation against the AI engine vendor. This is because the AI engine vendor would be exposed to both statutory damages and ongoing damages based on its future profits, and would remain in danger of an injunction that could prevent further commercial use of the AI engine. In other words, the AI vendor's entire future business model built around the compromised AI engine could be at stake. But if the AI engine could unlearn the compromised content, the leverage in litigation could be reset, and the AI engine vendor could mitigate its ongoing and future risks and financial exposure. In this case, the AI vendor would still remain liable for past damages, but the risk of a crippling injunction that could suspend or terminate the AI vendor's business would be removed.

4. Detecting Compromised Data and AI Unlearning

Given the risk and benefits discussed above for both content owners and the AI industry, increasingly more efforts are being directed in the AI industry towards both detecting use of compromised data (an issue of primary concern to content owners and regulators) and unlearning (a solution driven by AI engine vendors and customers).

(a) One approach to unlearning could be to feed to the neural network a large data set that is not compromised and that generally overlaps in scope and output characteristics with the compromised dataset. But this would only dilute the effects of the compromised dataset, and complete unlearning would theoretically never be completely achieved.

(b) Another approach could be to completely retrain the AI engine without the compromised data. But that could be prohibitively expensive (e.g., some generative AI engines cost hundreds of Millions of Dollars to train, identifying a small set of compromised data in a huge input dataset could be very expensive and time consuming, etc.) or may be virtually impossible (e.g., second order engines trained based on outputs provided by compromised AI engines).

(c) To completely reverse the learning that occurred based on a compromised input dataset, the weights of the neural network would have to be modified such that the effects of the compromised data are completely removed. But this is challenging to achieve fully. For example, adding noise to the weights of a neural network or otherwise modifying it intentionally to undo the effects of compromised data can practically achieve forgetting (see https://arxiv.org/abs/2007.02923), but risks compromising the accuracy and performance of the AI engine. Another approach suggests "scrubbing"' the weights of the neural network clean of information about a particular set of training data, so that any probing function of the weights of the compromised AI engine is indistinguishable from the same function applied to the weights of a network trained without the data to be forgotten (see https://arxiv.org/abs/1911.04933).

(d) The AI industry realizes the existential threat posed by compromised data that is used on a wide scale and that propagates to compromise AI engines at multiple levels. Various lines of research into AI unlearning have been initiated, and this area will certainly grow fast. For example, Google and various academic institutions and researchers have organized a Machine Unlearning Challenge designed to encourage further research into this area (see https://unlearning-challenge.github.io/).

(e) A critical question for both AI engine vendors and content owners is how to determine whether compromised data was actually unlearned by a neural network. An effective probing method uses membership inference attacks (MIAs) to recognize differences in a compromised AI engine's predictions using the inputs on which it was trained versus other inputs by using adversarial machine learning (see https://arxiv.org/abs/1610.05820). Some bad news for AI engine vendors is that in some cases MIAs could be used to infer whether certain data was used to train a model even if the data was deleted from the training set database, and potentially to even extract such data (e.g., MIAs could be used to extract personal data that was used to train a neural network).

5. Arms Race between Content Owners and AI Vendors

We would expect that more tools will be developed over time to help protect content owners' rights by identifying use of compromised datasets to train AI engines. At the same time, we expect to see AI engines improve their ability to unlearn compromised information, and possibly also try to try to obscure the use of compromised data (e.g., by insulating against MIA probing).

An arms race between content owners and AI engine vendors will likely expand and continue for years to come. In this context, effective and verifiable AI unlearning could be a material development that would ultimately benefit both content owners and the AI industry.

6. Conclusion

We expect and hope that the AI industry will develop better tools to determine when compromised data was used to train AI engines, and effective unlearning methods for AI engines to unlearn compromised data.

Overall, effective unlearning of compromised training datasets would be a major positive development in the AI industry, and would instill additional confidence in the ability of the AI industry to comply with applicable laws and regulations (including consumer privacy laws) and to respect the rights of the content providers.

Effective unlearning will also help achieve an effective balance between the rights of content owners and AI engine vendors, which will in turn protect and bring more certainty to the use of AI by business customers and consumers.