ISSN 2281-5147

WHAT AI IS STEALING ! DATA PRIVACY RISKS IN AI

Soumia Zohra El Mestari*

Abstract

Even if we may not realize it, AI's presence in our lives is increasing at a great pace. Most technological services we use nowadays are driven by AI, and that could be good news since AI's aims to improve the quality of the services. Unfortunately, to work well, AI greedily feeds on user data: AI models collect, process, and store a great deal about us, which is a problem if such sensitive information is leaked. This chapter discusses that this risk of AI's leaking personal data is not only hypothetical and suggests how to mitigate it.

Table of Contents

* PhD student at the Sociotechnical Cybersecurity (IrisC Group) SnT, University of Luxembourg, Soumia.elmestari@uni.lu

Soumia Zohra El Mestari is a PhD student at the university of Luxembourg supervised by Pr. Gabriele Lenzini . Her research interests are in Machine Learning, Trust and Transparency in data-driven tools and Privacy-Preserving machine learning. Prior to joining the Sociotechnical Cybersecurity Interdisciplinary research group, IRiSC, headed by Prof. Gabriele Lenzini, Soumia worked as a machine learning engineer and data analyst. Currently she is pursuing her PhD in IRiSC funded by the interdisciplinary EU project Legality Attentive Data Scientists (LeADS). This work is supported by the European Union's funded project Legality Attentive Data Scientists (LeADS) under Grant Agreement no. 956562.

ISSN 2281-5147

Keywords

Privacy Preserving Machine Learning – Machine Learning – Membership inference attack – Artificial Intelligence – Privacy Enhancing Technologies

1. Introduction

Artificial intelligence (AI) is revolutionising how we perform certain tasks by making it more tech-dependent. Today, AI tools can perform sophisticated tasks more efficiently than we can. These tasks include, for example, video and audio processing [4], natural language understanding [1], summarising and generating content [8, 11], and decision making. However, today, most people are not even aware of the number of AI-based tools they use on a daily basis; since once the technology is spread and used by everyone, it is no longer referred to as AI but rather seen as a mainstream tool, for example, receiving new social media content based on your interest is seen as a "the normal way" by which any social media feed operates. This integration of AI in life often makes users forget that AI is involved in the first place.

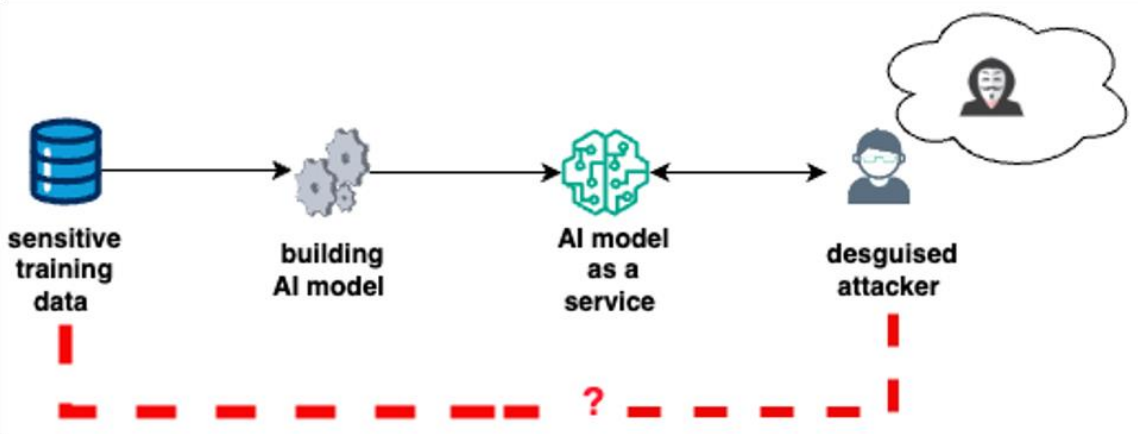
AI tools have the ability to enhance their performance by learning from feed- back and data collected from their environment. Thus, the word "intelligence" comes from this particularity of learning from data without explicit instructions on how to solve the tasks at hand. In this context, when we mention data, we mean huge volumes of

ISSN 2281-5147

images, texts, videos, audio, search feeds, in some cases, health records, social media preferences and anything that can be recorded and stored by electronic means.

The closer the purpose of the AI tool is to humans, the more sensitive the data becomes. For example, an artificial intelligence model that helps doctors make diagnoses needs to be built using patients' health records which are considered highly sensitive. Similarly, targeting a specific range of users for online ads calls for the collection of many users' online traces, including their search history. From another angle, the more data we feed to these tools, the better they become. This data-greedy nature can be seen in the viral ChatGPT-3, which has been trained with roughly 45 terabytes (TB) of text, nearly a trillion words [2]; and despite the huge amount of data collected to build ChatGPT, this AI model continues to collect more data to improve its performance on a regular basis.

This wide adoption has been a game changer in many fields, promising increased efficiency and enhanced cost-over-convenience ratios. However, like any other technology, AI has drawbacks: its efficiency and proficiency come with a heavy privacy bill. AI-tools feed on data, requiring more input data to deliver more accurate predictions, but this data hunger is a threat to our privacy.



2. All emerging technologies raise privacy issues

The growing need for more data to build efficient AI models made regulators and ethicists run the marathon in an attempt to prevent the misuse of data in AI by setting out data protection regulations that establish what should be done so that everyone's data are used properly and fairly. However, the evolution of artificially intelligent techniques is faster than the ability to regulate them, which opens a huge gap in data

ISSN 2281-5147

protection regulations when AI is involved in the process. To explain, although current data protection laws attempt to address these issues, the way AI tools process data is more complex, and regulations are not always flexible to address these complexities. Furthermore, regulations emphasise principles like fairness and privacy without a clear guidance on how to achieve them technically.

There is an inherent opacity in the way AI models operate and specifically in their learning processes, so understanding what steps are taken by these tools to produce certain decisions is challenging [10]. This opacity complicates attempts to detect and prevent data leaks. Under the hood, scientists still face the struggle to justify or even explain the learning patterns that govern the decision-making of a given AI model, which is not good news for lawyers and policy makers.

A significant distinction between AI and traditional analytics technologies is the ability to automate many tasks that humans used to have control over, such as data storage, processing, and maintenance. Thus, AI can interfere by modifying and automating the current applications in which data are used and consequently affects the privacy implications of these applications with little to no accountability to be redeemed. For example, using CCTV cameras for public surveillance is fairly common in contemporary society. This role was previously performed by security staff and in the first implementations of CCTV cameras, the task of refining and analysing the clips of videos was still performed mostly by a human staff and generally done in case of a security-threatening event. However, nowadays, when this technology is coupled with facial recognition software, a camera network could become a much more invasive privacy tool. Thus, the issue is in the way data are processed and the opacity that covers this processing; traditionally, when these tasks were mainly performed by humans, the risks and leakages were simpler to detect, and the responsible for the harm was easier to point out and hold accountable. However, with modern tools that include different layers of AI and little knowledge about processing details, it is harder to spot privacy issues, prevent them, or identify who is responsible for ensuring that they are held accountable.

Furthermore, building AI models to mimic human behaviour, such as voice- based conversational abilities, can give the illusion that these tools have human characteristics. Consequently, we find that users would deal with it as if it were a human being [9], thus giving away sensitive personal data without realising that these personal data may be processed in ways that may not be in line with one's perception of privacy.

ISSN 2281-5147

2.1 How do these challenges look like at the engineering level?

Under the hood, AI models can unexpectedly leak data. First of all, AI models offer results and outputs based on user inputs, for instance, the social media feed is full of ads for products that match the interests of the user since the underlying AI models for this purpose have already received a considerate flow of data that include, but not limited to, browsing history, previous interactions (Likes, comments, or watching video clips) on products or topics that are in line with the user's preferences, adding to that even the messages he may have exchanged with other persons that express these needs or even the interest of these people who he frequently interacts with assuming that socially close people generally share similar interests. All these inputs are sent to servers where AI models are deployed, in other words, data are stored on servers that users ignore their locations with no tangible guarantees to prove that entities who govern these servers will not use the data for other purposes or even monetise them.

In addition to the amounts of data that users have to disclose to the AI services to obtain results like predictions, recommendations, etc. AI is also built from data that are generally collected under a set of terms that do not include the free disclosure of these data to the public. In the best scenarios, user consent is requested for the purpose of building or maintaining AI models not to put the data in public for anyone to see and process. The bad news is that AI models can still leak these data in such a way that anyone using an online service that makes use of AI can push the AI model to leak the data that were used to build it. To picture the seriousness of this potential risk, one may agree on the usage of his medical data to build models that may help the diagnosis of similar pathologies he suffers from in an attempt to contribute to rapid recovery and improving medical services for a larger public. The data sharing terms here include only using the data to build an AI that helps diagnosis, however, when this AI is used to leak information to the large public, this may cause serious social, economic, and legal implications. This leakage can be on different scales from revealing that a certain person participated in a given study [7] to an actual recovery of their entire data [12] [3].

If the challenges mentioned above do not shake our awareness about the risks we face, then the following will certainly do!

ISSN 2281-5147

2.2 The privacy enemies may get away with it!

From a different perspective, identifying privacy issues in machine learning sys- tems can be challenging. In most scenarios, the threat is not detected until a significant data breach occurs! More critically, detecting a potential attacker who shows no signs of malicious behaviour remains a difficult problem. In other words, many attacks that may be conducted against the AI models cause one major problem: They can go undetected, and it is hard to hold the enemy in the loop (i.e., the attacker) accountable for the privacy breach they caused.

The attacker will use the system the same as any benign user who uses it, yet still leak the AI model data, which not only enlarges the pool of potential attackers but also decreases the chances of proving their malicious behaviours, thus holding them accountable and preventing their actions.

3. Are there any solutions offered to mitigate these risks ?

Privacy Enhancing Technologies (PETs) are frequently suggested as means of protecting personal data and achieving general trustworthiness according to cur- rent EU regulations on data protection and AI, this trustworthiness is important to ensure a safe usage of data for the best benefit of society. This set of tools is generally promoted as a means of achieving PPAI (Privacy Preserving AI), also known as PPML (Privacy Preserving Machine Learning)[1]. PETs offer privacy guarantees that depend on how they are applied. Different PETs offer different privacy guarantees and defend against different privacy risks, and there is no Privacy Enhancing Technology (PET) that can solve all privacy issues for a given AI system. Thus an off the shelf usage of these tools is not sufficient to render a privacy invasive AI tool into a perfect privacy-friendly AI. Switching the vision by placing these PETs under a legal lens makes the situation more confusing. The appropriate measures to be used to ensure legal compliance for an AI tool must be built on solid grounds, including an analysis of the whole data flow against the legal and technical guarantees that this flow must ensure.

[1]Machine Learning is a subfield of AI, both terms AI and machine learning can sometimes be used interchangeably

ISSN 2281-5147

4. Research Questions, Findings and Limitations

The first aspect of this project involved exploring the relationship between the requirements outlined in the EU data protection regulations and the actual privacy risks.

This includes creating detailed threat models[2] that take into account the stakeholders involved, the infrastructure where AI is deployed, and various stages of the process before and after implementing PETs.

This kind of analysis takes into account the trust relationships between the different stakeholders along with the guarantees that PETs are designed to offer a direct method of establishing a comparison between the desired privacy guarantees and those actually achieved [6].

This model may look complex, but building an AI tool and deploying it as a system include the participation of many entities depending on the system design, these entities may have different trust assumptions among themselves. For example, the user can trust that the entity deploying the AI model will actually process the data in a correct way to give the actual desired output to the user without outputing a wrong result. The same user may not trust this same entity to keep a copy of their data, in this particular example the PETs used must be really studied to satisfy the privacy guarantees of each entity without compromising neither the correctness of the output nor the privacy guarantees of the other entities.

It is important to recognise that while PET offer important protections, they also have limitations and disadvantages that need to be clearly communicated and considered when aiming for legal compliance in this field.

One of the key findings of our analysis was the shortcomings of current regulations in addressing certain complex AI scenarios, such as when AI models are repurposed, a practice known as transfer learning. This practice can potentially undermine the principle of allowing users to have control over their data. Transfer learning involves changing the purpose of an AI tool to perform another task. This can be achieved even without reusing the original data, making it difficult to detect or prevent data leakage. Legal mechanisms like informed consent struggle to keep up with the various

[2] In security analysis a threat model is a practice that studies a given system by identifying the parties that must be threatened and the potential threatening parties along with the threat points which symbolises the points of interactions between stakeholders that may constitute a risk on one another.

ISSN 2281-5147

potential transfer learn- ing purposes that can arise after a AI model has been built, posing challenges in informing users comprehensively about the data processing objectives. The legal examination of this issue delves into uncharted territories and contentious issues, including the issue of AI model ownership and whether people can claim co-ownership of a model developed with their data. This analysis highlights the inadequacies of existing EU legal tools to address complex AI issues and the limited adaptability of data protection regulations in addressing technical AI challenges [5].

The problem of safeguarding the privacy of the data has an interdisciplinary nature. AI is now being used by organisations and large tech companies, placing a great responsibility on engineers and decision makers to comply with data protection regulations. However, tailoring the technical implementations to the legal provisions faces many drawbacks.

In this type of techno-legal issues, one of the main challenges lies in the terminology used and how the casual use of terms like "anonymisation" can be misleading. In the EU data protection regulations, data that have been anonymised do not meet the criteria to be classified as personal data, which means that the processing of these data is not subject to the same restrictions under the GDPR (General Data Protection Regulation).To this end, a study to explore the idea of anonymisation, which goes beyond mere technical aspects, is crucial. Data anonymisation is like putting on a disguise for sensitive information. Imagine that a data-holding entity has a list of names, social security numbers, and addresses stored in a database. Anonymisation ensures that even if someone gets hold of these data, they will not be able to directly connect them back to specific individuals. The goal is to protect people's privacy while still allowing useful data to be shared and analysed. Unfortunately, the oversimpli- fied link between the legal definition of anonymisation and the technical tools called 'anonymisation algorithms' often leads to their limited use. The term 'anonymisation' as defined in the regulations can be mapped to many PETs including a kind of PETs that is also known under the name of 'anonymisation techniques,' which creates confusion for engineers who may confuse the legal definition with the technical one assuming that the regulation refers only to the set of tools known as 'anonymisation algorithms.' The tricky point here is that in certain scenarios the anonymisation algorithms are insufficient to satisfy the legal provisions and thus result in an underestimation of privacy risks and of the PETs guarantees by both the regulators and the engineers. This superficial approach may hinder effective data processing and

ISSN 2281-5147

also cause confusion within the tech community. Such misunderstandings can harm stakeholders, especially those meant to be protected by regulations.

To put the reader in the view, we describe the leakage risks and the challenges in detecting the risk and defend against it via a simplified example of a secret spy game.

4.1 The Secret Spy Game: Membership Inference Attacks

Imagine your favourite puzzle: AI systems are like that, solving complex problems. But sometimes they accidentally reveal secrets to attackers. These attacks are like invisible ninjas. They do not shout, 'Hey, I am attacking!' Instead, they blend in with regular users. Membership inference attacks are no exception. Imagine playing a super cool spy game. But instead of chasing bad guys, you are trying to figure out secrets about a secret club. Here is how it works:

The Secret Club: Imagine that there is a secret club (let's call it AI Model'). This club knows how to do cool things like recognise cats, dogs, and even unicorns in pictures! But the club has a hidden secret: It was trained using special pictures (such as a secret recipe).

The Spy (Attacker): You are a spy! Your mission: find out if a specific picture was part of the secret training. For example, you want to know if a picture of your cat was used to train the club.

The Clues (Model Output): The club gives you clues. When you show it a picture, it says, 'I am pretty sure this is a cat' or 'Maybe it is a dog?' These clues are like secret messages from the club.

The Sneaky Trick (Membership Inference Attack): You use these clues to guess whether the picture was part of the secret training. If you are right, you have cracked the code! You know whether your cat's picture was in the secret training of the club.

Why does it Matters: Imagine if the club were trained on medical records. Identifying which records were used, you could guess someone's health condition! It is like saying, 'Hey, this person's medical information was part of the secret training; maybe they have a unicorn allergy!' The Challenge The spy game is tough because you do not get to see the secret training pictures directly. You only get the club's hints. Remember, membership-inference attacks are like playing detective with AI models

ISSN 2281-5147

4.2 Catching the spy in the jungle

Spotting your attackers before they leak the data of your model to the large public is difficult and can be impossible in some cases, so imagine looking for them in complex scenarios like the scenario of repurposing AI models or when federated learning[3] is involved ?

The repurposing use case is studied in the context of those large AI tools known as language models. Language models like ChatGPT are making their way into our daily lives in an invasive way. The amount of privacy leakage that these models have proved is alarming. For example, attackers can trick ChatGPT or any other AI tool that generates text to reveal sensitive information about someone's data that were used to build this model by asking it precise questions about this individual such as to reveal their phone number. Thus, putting them under another test in a more complex setting such that of transfer learning have shown interesting results, and despite the general belief that the practice of re-purposing in its technicalities helps in preserving the privacy of data and making the mission of the spies (the attackers) more difficult, when the AI model is a language tool the game balances change.

These AI language tools can become a spy helper disguised as your confident writing assistant.

4.3 Not only that the spy may be a member of the privacy team!

One of the PETs that has a good reputation in the privacy team is a technique known as Federated Learning. In Federated Learning, user data never leave the user's device, and AI tools are created so that multiple users contribute to the construction of the building blocks of AI tools under the governance of one entity called the aggregator. The privacy angel called the aggregator does not have access to the data of the users and its job is to assemble the building blocks that are sent by all users to build the AI model; these building blocks are, however, built by the client from their data.

One of our most interesting studies showed how this aggregator may modify the way users build their AI building blocks to further use those building blocks to extract users' data, the power of this entity being an aggregator is not only underestimated, but when the entity is clearly doing a malicious behaviour it goes undetected! Our

[3] Federated learning is a technique to build ai models by using data from different entities without having to merge all their data in one place, so the training happens in a collaborative way where each data holder does a portion of the processing

ISSN 2281-5147

results suggest that aggregators can spot which users to target and perfectly push them to reveal their data without the users spotting this behaviour until it is already too late and data have been revealed.

Our study also explored the different defence mechanisms and the limita- tions of each. The performant defence strategy included periodically testing aggregators, and once privacy-invasive behaviour is detected, users should opt out of the collaborative learning process (federated learning process).

4.4 Exploring the horizons!

Despite the wave of research efforts in tailoring the privacy risks of AI, the limitations and challenges in the field are severe.

Privacy-preserving AI methods often involve adding noise or altering data to protect privacy. But this can affect the AI tools' performance. Think of it as baking cookies: If you add too much flour (privacy protection), the cookies might taste bland (low accuracy). If you add too little, they might fall apart (privacy breach). Researchers are working hard to find the right balance between privacy and accuracy, but it is a delicate dance.

In addition, implementing privacy-enhancing techniques requires expertise. It is like assembling a puzzle with many pieces. Developers need to understand how to set privacy parameters, choose the right tools, and ensure that the model does not accidentally leak sensitive information. It's a bit like building a sand-castle: You need to know where to place each grain of sand to keep it sturdy and safe. In summary, privacy-preserving ML is like protecting a secret recipe. You want to share the delicious cookies (ML predictions) without revealing all the ingredients (private data). Finding that sweet spot between privacy and accuracy is the challenge!

5. Conclusion

This project aims to study the privacy risks of using AI without being aware of its risks. Allowing your data to circulate without being aware of the different ways your data may be exposed, manipulated, and shared. Our findings show that the risks are hard to detect and can go without being noticed. In addition to that, and from a technological perspective, privacy enhancing techniques are still immature to be used in an efficient way, thus the researchers are still trying to enhance the privacy enhancing versions of AI to achieve the same service quality results without

ISSN 2281-5147

compromising the privacy of the users. Furthermore, regulations need to gain more flexibility to capture all risks and provide users with the necessary legal protection they need.

References

[1] Md Ali, Nawab Yousuf, Md Rahman, Jyotismita Chaki, Nilanjan Dey, KC Santosh, et al. Machine translation using deep learning for universal networking language based on their structure. *International Journal of Machine Learning and Cybernetics*, 12(8):2365–2376, 2021.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Ka- plan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Nicholas Carlini, Florian Tram`er, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, U´lfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021.

[4] Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.

[5] Soumia Zohra El Mestari, Fatma Su¨meyra Do˘gan, and Wilhelmina Maria Botes. Technical and legal aspects relating to the (re)use of health data when repurposing machine learning models in the eu. In Stefan Schiffner, S´ebastien Ziegler, and Meiko Jensen, editors, *Privacy Sympo- sium 2023*, pages 33–48, Cham, 2023. Springer International Publishing.

[6] Soumia Zohra El Mestari, Gabriele Lenzini, and Huseyin Demirci. Pre- serving data privacy in machine learning systems. *Computers & Security*, 137:103605, 2024.

[7] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s), sep 2022.

ISSN 2281-5147

[8] Touseef Iqbal and Shaima Qureshi. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528, 2022.

[9] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630, 2021.

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligenæ*, 1(5):206–215, 2019.

[11] Liuyang Wang and Yangxin Yu. Research on text summarization generator method based on input text linguistic features and copy mechanism. In *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 586–590. IEEE, 2021.

[12] Xi Wu, Matt Fredrikson, Somesh Jha, and Jeffrey F. Naughton. A method- ology for formalizing model-inversion attacks. *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370, 2016.

ISSN 2281-5147