https://medium.com/@RocketMeUpCybersecurity/privacy-preserving-machine-learning-how-to-train-models-without-compromising-data-866a825097d2
https://www.smarsh.com/blog/thought-leadership/generative-AI-and-data-privacy-the-challenge-of-PII-use-in-training-data-sets
https://www.zscaler.com/blogs/product-insights/ai-cybersecurity-navigating-gdpr-privacy-laws-and-risk-management

The three main factors are: AI data volume needs, individual privacy rights, and data privacy laws.

- AI needs a lot of data to work with, and organizations often choose to make PII a part of the training set for the AI. This choice is at the very least not very far off from a data breach, thus making this a major issue. There are also other current standards that this may violate, including the principles of purpose limitation and data minimization.
- Under many current data privacy laws, individuals have the right to have their PII changed, accessed, and deleted from the storage of any organization in possession of such data. It appears however that once PII is used in training an AI model, it is remarkably difficult, if not impossible, to remove that and only that data point. Even if this is done, it is likely that the AI model's performance will degrade as a result of that removal. Some of the laws that govern these standards are: EU's GDPR, California's CCPA/CPRA, and even HIPAA.

These problems are solved at least in part by privacy preserving machine learning (PPML), which aims to essentially facilitate the training of these AI models without compromising data security and privacy. Some of the key principles are:

- Data confidentiality
    - Encrypt or anonymize (or even redact) all data used to train an AI model.
    - Minimize the amount of PII used in training to only what is strictly necessary.
- Model security
    - Models must be made to resist reverse engineering so as to protect all data, especially PII, from a breach.
    - Models must remain effective in the presence of an attack to resist breaches.
- Compliance and Ethics
    - Comply with data privacy laws, such as GDPR, HIPAA, CCPA/CPRA, or others.
    - Data must always be handled through proper channels to ensure security. How a company handles its users' data will directly affect if and how much its users trust the company.

Techniques:

- Differential privacy
    - Add controlled noise to model output or data so that an AI model can be helpful to the user without exposing PII.
    - Alternatively, a privacy budget can be adhered to, which limits how much information about an individual can be revealed. This entails more risk, and proper value must be placed on each data point.

- Federated Learning: train models locally, and only share model updates rather than transferring all sensitive data to a central server.
- Homomorphic encryption: allows computations to be performed on encrypted data without needing to decrypt it.
- Secure Multi-Party Computation (SMPC): multiple parties allowed to jointly compute a function over their inputs while keeping those inputs private. Collaboration and security.

Some of these components may increase model complexity, and some techniques like homomorphic encryption and SMPC may introduce significant computational overhead.

Overall, it appears that the main challenges with managing PII ethically in AI models relate to how the data is used in AI models. Since they use massive data sets, companies often train the models using PII they have, however this presents a huge security risk. It's not easy to fix this though, as specific data points can't always be targeted and removed from the AI's data pool without causing performance degradation. This makes compliance with data privacy law tricky.

The principles of PPML are going to be very necessary to lean on for our policy decisions. These ideas outline how to build an AI model from the ground up so that it is both effective and compliant. We will need to focus on how to change what already exists, but it would be very good to focus on how models should be trained for the future.

Implications of GenAI and the new data privacy laws:

Difficulty in isolating and removing specific PII: GenAI typically is trained on massive datasets. This data is also distributed quite vastly, and thus it can be at best extremely difficult to identify and remove specific requested data from the model's dataset.

Potential model degradation: Depending on how significant the removed datapoint is, a successfully removed datapoint might vastly change how an AI responds to queries.

Compliance: Failure to comply with the request to remove data is possibly a violation of data privacy law, which carries significant repercussions for the organization.

Synthetic data and differential privacy might help in training a model without exposing sensitive information. Documentation and data lineage might help to identify where PII is used in a training set to ensure easier removal.

If a specific data subject requests deletion of their PII and that PII was used for generative AI content creation, would all the associated content generation output from that model also need to be deleted and the model retrained with the new training set?

Realistically, there is no way that all GenAI content using specific data points could be deleted because there is simply no way to track all of that content on the internet. Once it's

there, it's there forever, so it isn't worth worrying about. The better course of action here is prevention.

AI is also hard to train with PII under GDPR because the law requires organizations to get explicit consent from the user whom that data belongs to before collecting that data for any reason.

GDPR also requires transparency and accountability, however many of these AI models are the opposite of transparent thus accountability is very hard.

Some solutions:

- Privacy first: anonymize, pseudonymize, and/or encrypt the data used in the AI training set to reduce the amount of PII that comes through to the surface. Thus the AI can still use some PII (in compliance with GDPR regulations) in its data set when absolutely necessary, but in other cases, redacted data can be used in place of where PII would normally be used.
- Explainable AI (XAI): implement AI whose decisions and data processing methods can be clearly articulated, demonstrated, and tracked. This would help to facilitate transparency in closer compliance with GDPR requirements.
- Rebuilding for compliance: AI models need to be built from the ground up in adherence to GDPR requirements, including implementing data loss prevention (DLP).
- Zero Trust: only authorized personnel get access to IT resources and data, and they are only given the minimum amount of access necessary to do their jobs for the moment they need it.

GenAI policy suggestions:

1. The use of PII in training Generative AI models
    a. Personally Identifiable Information (PII) shall not be used in the training of GenAI models (training sets) except if the company training the model can produce a clear and written request and reason for the use of PII in their training and submit it to the governing policy body for approval.
        i. Reasoning for using PII in training sets must include a clear explanation of why PII cannot be substituted by anonymized data for the desired use.
        ii. Submission of such a request is not guaranteed to be approved; it must first be approved by the governing body.
    b. In place of PII, anonymized data must be used.
        i. Anonymized data is data that has been altered or replaced in such a way in which it cannot be reversed back to the original data, no matter how much additional information is used. This data reassembles real data but is in fact fake, and is safe to use in an AI training set.

c. When PII is allowed to be used:
    i. The company must obtain written permission from the data subject in order to use that data. That document must explain in clear and plain wording what data would be used and what specifically it would be used for; if the document does not contain this type of wording, it is not legally valid. The company is bound by the obligation to only use the data requested and only for the specific reasons outlined in the agreement; violation of this will result in consequences for the company. Such contracts are also not allowed to request the subject to release the rights they have over their data that are outlined in any law applicable to the region of operation.
    ii. Pseudonymization should be used in order to help protect data being used in the training set. Pseudonymization is the technique of replacing PII fields with identifiers that can be traced back to the original data by use of a key list.
    iii. Encryption of the data set should be done in order to hinder the effectiveness of breaches. When possible to implement without performance degradation, homomorphic encryption should be used on PII in the data set.

2. Models should be built from the ground up with adherence to the concepts of Privacy-Preserving Machine Learning (PPML) and Explainable AI (XAI).
    a. PPML concepts include adherence to regional laws governing security, prevention of reverse engineering by securing the model, adversarial robustness, data confidentiality, and minimization of data use.
    b. XAI is about ensuring that an AI model's decisions and data processing methods can be clearly articulated, demonstrated, and tracked. This would help to facilitate transparency.
3. Models need to be implemented with the concept of zero trust.
    a. Zero Trust: only authorized personnel get access to IT resources and data, and they are only given the minimum amount of access necessary to do their jobs for the moment they need it.

https://www.nist.gov/news-events/news/2025/03/nist-finalizes-guidelines-evaluating-differential-privacy-guarantees-de
https://gdpr-info.eu/
https://www.geeksforgeeks.org/what-is-data-anonymization/