Mesterséges intelligencia-specifikus adatvédelmi incidensek

Európai értékek a legújabb technológiák között konferencia 2025.

Budapest, 2025. május 22.

Tóth Péter, ügyvéd
OPLgunnercooke

Az incidens fogalma

A biztonság olyan sérülése, amely a továbbított, tárolt vagy más módon kezelt személyes adatok véletlen vagy jogellenes megsemmisítését, elvesztését, megváltoztatását, jogosulatlan közlését vagy az azokhoz való jogosulatlan hozzáférést eredményezi.

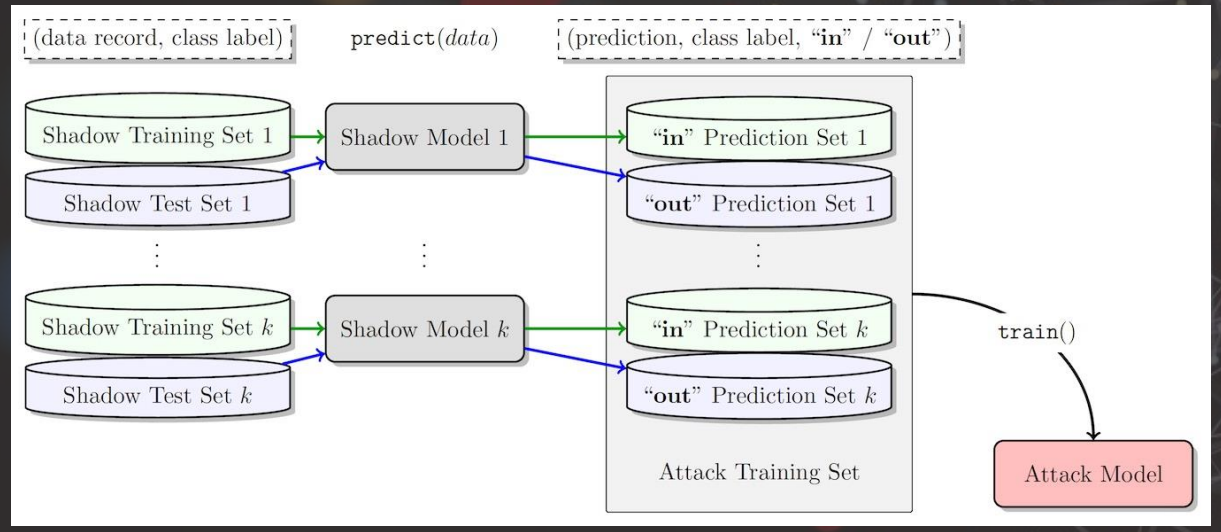
Főbb AI incidens típusok

Incidens neve	Mire irányul a támadás?	Milyen jogsértést okoz?
Membership inference	Annak meghatározása, hogy egy adott adat része-e a tréning adatszettnek	Adatvédelmi incidens
Model inversion	Tréning adatok kinyerése	Potenciális adatvédelmi incidens
Model evasion	Az adott modell megkerülése	Jogellenes adatkezelés / adatvédelmi szempontból nem értelmezhető
Model extraction	Az adott modell bizonyos részeinek kinyerése	Potenciális adatvédelmi incidens
Model poisoning	Az adott modell bizonyos jellemzőinek megváltoztatása	Jogellenes adatkezelés?
Data poisoning	Tréning adatot változtat meg a modell elrontása érdekében	Adatvédelmi incidens, jogellenes adatkezelés?
Property inference	Az adott modell bizonyos jellemzőinek kinyerése	Potenciális adatvédelmi incidens

Data poisoning, Model poisoning

- *Egy sztenderd biztonsági incidens, amely AI tanító adatra vagy a modellre irányul*
- *Probléma: a modell innentől kezdve nem úgy működik, ahogy azt tervezték; jogellenes-e így az adatkezelés?*

Membership inference



- *Annak kihasználása, hogy a modellek a tanító adataikra jobban reagálnak*
- *Probléma megjelenése: MLaaS – Google, Amazon*
- *Felépítés: White-box, black-box*
- *Probléma: akkor működik, ha már egyébként is megvan a hozzáférés az alap adathoz, ez csak „további” személyes adat megszerzésére alkalmas. Elég szigorú-e ehhez a GDPR?*

Kép forrása: R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017, Membership Inference Attacks against Machine Learning Models

Model inversion

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

- *Alapja, hogy bizonyos esetekben a tréning adatok „nyomot hagynak” a modelleken.*
- *Tipikus esetek: arcfelismerés, egészségügyi adatok*
- *Hagyományos megoldások, GAN, LLM-re optimalizált modellek*
- *Probléma: nagyon minimális információval is működik, kevés a hagyományos értelemben vett biztonsági sérülés, nem a kérdéses adathoz férünk hozzá, hanem „rekreáljuk” az adatot. Incidens-e ez?*

Kép forrása: J. Zhang, C. Li, J. Ye, and G. Qu. 2020. Privacy threats and protection in machine learning.

Biztonsági kockázatok jogszabályi lefedettség nélkül?

Köszönöm a figyelmet!

Tóth Péter, ügyvéd, legal engineer
peter.toth@gunnercooke.com
OPLgunnercooke