

Neoprávněné užití díla při trénování modelu generativní AI

Případová studie

ČPIT 2026

10/9/2026

| Data, dataset, model

- **Data** (raw data): *mohou obsahovat / představovat dílo? ANO*
↓ (výběr, třídění, čištění, kontrola)
- **Dataset** (obecný vs. specializovaný, textový vs. obrazový) *může obsahovat dílo? ANO*
↓ (strojové učení)
- **Model** (obecný vs. specializovaný...)
 - *může obsahovat kopii díla? NE (jde jen o matematický model)*
 - *může představovat jiné vyjádření díla? ASI ANO (memorizace)*

| Memorizace

= jev, kdy model doslova nebo téměř doslova reprodukuje / opakuje části tréninkových dat namísto toho, aby z nich pouze odvozoval obecné vzorce.

Memorizace se projeví ve výstupu modelu – tj. tehdy, když model reprodukuje části tréninkových dat.

- Memorizace **vzniká při tréninku** AI modelu
- Memorizace **existuje (přetrvává) uvnitř AI modelu**, jako určitý stav parametrů.
- **Poznat** ji však lze **jen navenek (na výstupu)**, tj. podle toho, co model generuje.

| Důsledky Memorizace

1) Prokazuje, že určitá data (např. dílo) byla užita při tréninku modelu:

Pro tyto účely se nemusí data výstupu projevit doslovně, stačí výrazná shoda s originálem („nepravděpodobné shody“, unikátní řetězce – *canary strings*, rekonstrukční útoky – *pokračuj v textu...*)

2) ASI zakládá odpovědnost za provozování systému AI založeném na tomto modelu:

I když je provozovatel AI systému odlišný od osoby, která AI model vytrénovala, může být tento provozovatel odpovědný za užití děl “memorizovaných” v modelu (i když sám díla při trénování modelu nepoužil)

3) Neoprávněné užití díla ve výstupu:

Pokud je shoda vysoká (v podstatě úplná shoda v rozsahu, který je postačující pro autorskoprávní ochranu), tj. ve výstupu je kopie chráněného díla => neoprávněné užití ve výstupu

| Lze Memorizaci zabránit?

1) Zabránit memorizaci na při tréninku modelu: LZE

Existují techniky, které tomu mají zabránit (např. *Goldfish loss*), nicméně takto vytrénovaných modelů je zatím málo (*Apertus*); většinou na úkor kvality modelu.

2) Zabránit tomu, aby se memorizace projevila ve výstupu: LZE

Technicky lze riziko, že se memorizace projeví ve výstupu, výrazně omezit až vyloučit.

Neoprávněné užití při trénování modelu a provozování AI systému

- český nositel práv
- při **trénování modelu** AI byl použit jeho obsah (data)
- **v modelu je obsah** memorizován
- Obsah chráněn autorským právem (autorské dílo) nebo právem souvisejícím (zvukový záznam, obrazový záznam, tisková publikace)
- rozhodl se podat žalobu
- chce ji podat u soudu v ČR, pokud možno aby byla posuzována podle českého práva

Má český soud pravomoc a příslušnost?

“provozovatel” a “tvůrce modelu” jsou dvě různé osoby

- Žalovaný č. 1: **tvůrce modelu**, který AI model vytrénoval
- Žalovaný č. 2 : **provozovatel AI systému** (služby) založeného na AI modelu
 - je z EU (např. Irsko)
 - **Nařízení Brusel I bis** (1215/2012)
 - zvláštní místní příslušnost (čl. 7 odst. 2) → místo škodní události
 - není z EU (např. USA)
 - **občanský soudní řád** (čl. 6 Brusel I bis)
 - § 87 písm. b) o.s.ř. -> místo (zjednodušeně), kde došlo k újmě (*Hejduk C-441/13*)

Služba dostupná v ČR => (snad) založena příslušnost v obou případech (zejména v případě memorizace na výstupu)

Aplikuje soud české právo?

- Bernská úmluva, nař. Řím II (864/2007), čl. 8 odst.1 - *lex loci protectionis*
- Potenciální zásah do autorských práv či práv souvisejících
 - -> právo země, pro kterou je uplatňována ochrana práv
 - **žalovaný č. 1:** užití v rámci tréninku modelu: užito mimo ČR, pravděpodobně v USA => neoprávněné užití v USA => **právo USA**
 - *Getty Images vs. Stability AI (UK), Gema vs. SUNO* (Německý soud aplikuje US právo)
 - **žalovaný č. 2:** provozování AI systému s memorizovaným obsahem: širší koncept sdělování díla veřejnosti – tam, kde je služba dostupná => **právo ČR**
 - *GEMA vs. OpenAI (Německo), Gema vs. SUNO*
 - Platí zejména pokud se v ČR projeví důsledky memorizace (tj. ve výstupu českému uživateli)

Došlo k užití díla?

- Při **trénování modelu (na vstupu)**: **ANO**, rozmnožování (v rámci datasetu atp.)
- Při **provozování AI systému (modelu)**:
 - Projevuje se memorizace?
 - ANO => **MOŽNÁ (spíše ANO)**
 - *GEMA vs. Open AI*: ANO (memorizace v modelu = užití díla v modelu i když nejde o přesnou kopii)
 - *Getty Images vs. Stability AI*: NE (v modelu nejsou uloženy kopie díla)
 - NE => **asi NE**, nedochází k užití, v modelu nejsou kopie
 - paradox: model je stejný, ale jen podle toho, jak je zamezeno tomu, aby se memorizace projevila ve výstupu -> různý závěr ohledně užití díla při provozování modelu
- Při **generování výstupu**:
 - **ANO**, pokud je dílo reprodukováno ve výstupu (*vyšší standard než memorizace*)

Bylo takové užití neoprávněné?

- Obecně: užití díla bez souhlasu (licence) = neoprávněné užití
- Užití při **trénování modelu (na vstupu)**:
 - Nejspíš cizí právo (USA)
 - Posouzení aplikace *fair use*
- Při **provozování AI systému (modelu)**:
 - České právo
 - TDM výjimka; Je možné TDM výjimku uplatnit na trénování genAI?*
 - Projevuje se memorizace?
 - ANO => **TDM výjimka se spíše neuplatní**
 - *GEMA vs. Open AI*: (TDM výjimka se neuplatní na užití, které jde nad rámec analýzy (tj. na memorizaci a reprodukci)
 - NE => **TDM výjimka se spíše uplatní** (*Kneschke vs. LION (DE)*, *DPG Media vs. HowardsHome (NL)*, *předběžná otázka k SDEU C-250/25*)
- Při **generování výstupu**:
 - České právo
 - Neoprávněné sdělování díla veřejnosti; odpovědný je provozvoatel AI systému (ne uživatel)

Jak prokázat (neoprávněné) užití?

- **Přímou kontrolou datasetu**
 - pokud je známo, jaký byl užit a
 - pokud je transparentní
- **Informace o datech / datasetu**
 - AI Akt (čl. 53(1)(d) + šablona pro veřejný souhrn obsahu, nelze zjistit:
 - zda byla použita konkrétní kniha, článek, fotografie
 - zda byla použita konkrétní webová stránka, pokud se nejedná o největší domény
- **Vyžádat si informace** – právo na informace podle směrnice 2004/48/ES
 - omezené možnosti pro prokázání, že došlo k neoprávněnému užití, spíše k prokázání rozsahu takového užití
- **USA: e-discovery** - je možné použít veřejně dostupné informace před soudem v ČR?
 - teoreticky možné, ale v praxi velmi omezené (*protective orders, GDPR*)
- **Memorizace (ale pozor na RAG!!!)**

RAG = technika, která umožňuje AI čerpat informace z externích dokumentů a databází (internet) namísto spoléhání se pouze svůj model => při použití RAG shoda ve výstupu nemusí prokazovat memorizaci

Čeho se domáhat?

- **Náhrada škody (ušlý licenční zisk / vydání bezdůvodného obohacení)**
 - Zpravidla fikce: až dvojnásobek obvyklé licenční odměny
 - Jaká je obvyklá licenční odměna
 - 3000 USD / knihu – *Bartz vs. Anthropic (US), settlement*
 - Sazebníky kolektivních správců?
 - Licenční smlouvy třetích stran v obdobných případech
 - Rozsah užití - právo na informace (směrnice 2004/48/ES + § 3 zákona č. 221/2006 Sb.)
- **Odstraňovací nárok**
 - „Odtrénování“ modelu AI?
 - V praxi nejspíše nemožné
 - Znepřístupněné AI služby v ČR? (teritorialita aut. práva – asi by tím formálně bylo splněno)

=> cílem zpravidla je „donutit“ provozovatele AI systému k uzavření licenční smlouvy

| Děkuji za pozornost!

Jiří Čermák

cermak@iplaw.cz
cermak@itpravo.cz