

Možnosti obrany online vydavatelů proti vytěžování jejich obsahu umělou inteligencí

Sdružení pro internetový rozvoj (SPIR) zastupující české online vydavatele zveřejnilo dne 7. července 2023 na svých webových stránkách¹ návrh standardizované komunikace, pomocí které se provozovatelé internetových stránek v České republice mohli automatizovaně vymezit proti vytěžování jejich obsahu umělou inteligencí. Návrh vznikl jako výsledek samoregulační aktivity a spolupráce podnikatelů v oblasti digitálních služeb a tvůrců obsahu, jejichž zájmy SPIR zastupuje.

Návrh standardizované komunikace, pomocí které mohli provozovatelé vyjádřit tzv. opt-out, vypracoval SPIR v souvislosti s nástupem umělé inteligence. Cílem doporučení z roku 2023 bylo nastavit transparentnější a předvídatelnější obchodní prostředí pro vývojáře umělé inteligence, provozovatele internetových stránek i samotné autory. V době přijetí výše uvedeného doporučení byly předmětem zájmu primárně tréninková data, která sloužila k vývoji obecných modelů AI.

K vytěžování dat dochází celoevropsky i celosvětově a do popředí zájmu se mimo tréninková a učící data dostávají také data využívaná k tzv. real-time odpovědím (asistenti umělé inteligence, online sumarizace a obdobné funkce). Od roku 2023 rovněž došlo k přirozenému vývoji nástrojů a způsobů, kterými mohou provozovatelé internetových stránek vyjádřit svůj postoj k automatizovanému vytěžování textů a dat (TDM). SPIR se proto ve spolupráci s Asociací online vydavatelů (AOV), Českou unií vydavatelů (ČUV) a Správcem licenčních práv vydavatelů (SLPV) z výše uvedených důvodů rozhodl aktualizovat jednotný standard, díky kterému bude snadné určit, zda a v jaké konkrétní podobě provozovatel internetových stránek opt-out pro TDM nastavil.

Tento jednotný standard využívá tzv. Robots exclusion Protocol (*robots.txt*), který pro další oblasti užití vznikl již před rokem 2000, formálně byl představen v roce 2019 jako IETF standard RFC 9309 a byl schválen v roce 2022. Standard umožňuje efektivně, transparentně a automaticky komunikovat s internetovými crawly tak, aby se provozovatelé internetových stránek mohli efektivně rozhodovat ve vztahu k různým způsobům a účelům užití a efektivně uplatňovat své právo podle čl. 4 odst. 3 Směrnice Evropského parlamentu a Rady (EU) 2019/790, případně obdobných zahraničních právních úprav (dále jen "výjimka TDM"). Tento standard je určen jako řešení pro vytěžování prostřednictvím internetových crawlerů, nikoliv pro jiné formy vytěžování dat.

Jednotný standard vypracovaný pracovní skupinou SPIR sestává z následujících dvou částí:

1. Užití dat k trénování obecných modelů AI (např. velké jazykové modely) a dalších AI systémů

Pokud si provozovatel internetové stránky nepřeje, aby jeho autorsky či obdobně chráněný obsah byl užíván:

a) pro účely trénování obecných modelů AI (na základě výjimky TDM), a

b) pro vytváření datasetů pro účely podle písm. a),

může na své internetové stránce aktualizovat soubor *robots.txt* o následující direktivy:

User-agent: MachineLearning

Disallow: /

Tato volba zahrnuje veškerý obsah dostupný na dané internetové stránce. Tato volba se nevztahuje na běžné internetové vyhledávání v rámci vyhledávačů, jejichž crawly nejsou tímto nastavením dotčeny, ledaže se jedná o AI módy těchto vyhledávačů (jako je např. AI Overview od Google).

¹ <https://www.spir.cz/online-vydavatele-se-vymezuji-proti-vytezovani-dat-umelou-inteligenci/>

2. Užití dat k trénování obecných modelů AI a dalších AI systémů a pro účely real-time odpovědi

Pokud si provozovatel internetové stránky nepřeje, aby obecné modely AI nebo systémy AI užívaly autorsky či obdobně chráněný obsah pro účely uvedené ad. 1 a zároveň pro účely tzv. real-time odpovědi (asistenti umělé inteligence, online sumarizace a obdobné funkce, typicky v rámci retrieval augmented generation) na základě výjimky TDM, může na své internetové stránce aktualizovat soubor *robots.txt* o následující direktivy:

User-agent: AI

Disallow: /

Tato volba zahrnuje veškerý obsah dostupný na dané internetové stránce. Tato volba se nevztahuje na běžné internetové vyhledávání v rámci vyhledávačů, jejichž crawlers nejsou tímto nastavením dotčeny, ledaže se jedná o AI módy těchto vyhledávačů (jako je např. AI Overview od Google).

Direktivy pro specifické User Agenty mají přednost před direktivami uvedenými pro User Agenta MachineLearning a User Agenta AI. Direktivy pro User Agenta MachineLearning mají přednost před direktivami pro User Agenta AI.

Provozovatel internetových stránek tímto nastavením jasně deklaruje opt-out z výjimky TDM a vůli jednat o kompenzaci za užití autorsky či obdobně chráněného obsahu s AI platformami a dalšími uživateli. Konkrétní nastavení pro konkrétní AI platformy může provozovatel internetových stránek následně nastavit (povolit / zakázat) prostřednictvím jejich specifických User Agentů (např. Apple-Extended; Google-Extended; Perplexity-User; Seznam-Extended; Open AI Crawlers atd.).

V návaznosti na výše uvedené je možné doplnit v souboru robots.txt také textový popis, který je pouze podpůrný a nastavení v bodu 1 a 2 mají přednost. Doporučujeme např. toto znění (či jeho anglický překlad):

Nastavení "User-agent: MachineLearning" a "User-agent: AI" jsou nástroje jednotného standardu SPIR pro automatické vytěžování textů a dat z této internetové stránky zejména ve smyslu čl. 4 směrnice 2019/790/EU.

Více o tomto standardu naleznete zde:

Kontakt pro případ jednání o získání práv

Tento standard neomezuje využití jiné formy zápočty pro deklaraci výjimky z TDM (např. textovou formou do patičky internetové stránky či využitím standardu [TDM Reservation Protocol apod.](#)). Upozorňujeme, že pro zabránění případným nejasnostem a rozporům je důležité, aby jiné formy zápočty byly využity způsobem, který bude souladný s nastavením tohoto standardu.