Wyobraź sobie, że jesteś artystą cyfrowym. Przez lata rozwijałeś swój charakterystyczny styl - unikalną paletę barw, kompozycję, kreskę. Pewnego dnia zauważasz, że popularny model sztucznej inteligencji generuje obrazy, które do złudzenia przypominają Twoje prace. Nie są to kopie, ale „coś" w nich jest znajome. Zaczynasz podejrzewać, że Twoje prace mogły posłużyć do trenowania tego modelu. Ale jak to sprawdzić?

Dziś nie istnieją narzędzia, które pozwalałyby twórcom, wydawcom czy instytucjom jednoznacznie ustalić, czy ich dane zostały wykorzystane do szkolenia dużych modeli generatywnych, takich jak ChatGPT, DALL·E czy Stable Diffusion. Wraz z rosnącą liczbą pozwów dotyczących praw autorskich (np. Getty Images vs. Stability AI czy New York Times vs. OpenAI), staje się jasne, że potrzebujemy narzędzi, które umożliwią audyt źródeł danych treningowych - także po fakcie, nawet jeśli dane nie zostały zapamiętane dosłownie.

Celem naszego projektu jest opracowanie skalowalnych i praktycznych metod pozwalających określić, czy - i w jakim stopniu - konkretne dane zostały użyte do trenowania dużego modelu generatywnego. Nasza centralna hipoteza brzmi: dane pozostawiają subtelne, statystyczne ślady w zachowaniu modelu. Jeśli umiemy je wzmocnić i odpowiednio zagregować, możemy uzyskać dowód na wykorzystanie naszych danych do trenowania modelu, nawet jeśli model nie odtwarza ich dosłownie.

W Zadaniu I opracujemy metody pozwalające nie tylko wykryć obecność danych, ale też oszacować, jaka część zbioru została wykorzystana. To ważne z punktu widzenia prawa autorskiego, które uwzględnia także „rozmiar i wagę" zapożyczenia. W Zadaniu II skupimy się na wzmacnianiu sygnału: opracujemy techniki pozwalające wykrywać wykorzystanie nawet małych kolekcji, takich jak prywatne portfolio artysty. W Zadaniu III zmierzymy się z problemem braku danych referencyjnych - czyli takich, co do których mamy pewność, że nie były użyte do treningu - i zaproponujemy sposoby ich syntetycznego generowania lub statystycznego uwzględniania różnic. Wreszcie, w Zadaniu IV rozszerzymy nasze metody na modele multimodalne, które uczą się równocześnie z tekstu i obrazu - czyli dokładnie takie, jakie coraz częściej spotykamy w nowoczesnych systemach AI.

Projekt ma nie tylko znaczenie naukowe, ale i społeczne. Wspiera właścicieli danych, umożliwia niezależny audyt i dostarcza narzędzi dla regulatorów oraz prawników. W dłuższej perspektywie może przyczynić się do bardziej etycznego i przejrzystego rozwoju sztucznej inteligencji. Wszystkie opracowane metody planujemy udostępnić jako otwarte oprogramowanie, by mogły być wykorzystywane w praktyce przez twórców, dziennikarzy, instytucje kultury czy organizacje zajmujące się ochroną danych.