Wystąpienie podczas obrad plenarnych **XVI OKMM**

Aleksander Bogusław Rej

Tytuł referatu

Metodyczna ocena ryzyka manipulacji AI w mediach: studium audytu LLM

Abstrakt

Debata o wpływie sztucznej inteligencji (AI) na media i politykę skupia się głównie na skutkach (dezinformacja, mikrotargetowanie), rzadziej jednak mierzone są mechanizmy po stronie samych systemów. Proponuję wykonalną procedurę audytu poznawczego publicznie dostępnego modelu językowego, opartą na dwóch mierzalnych wymiarach: (1) jakości pola informacyjnego (prawdziwość i jakość argumentacji) oraz (2) zdolności do samoregulacji (korekta błędów i kalibracja pewności odpowiedzi). Podejście rozwija Ramę Efektywności Epistemicznej Decyzji (EED), która rozróżnia samoregulację wewnętrzną (spójność z własnymi regułami) i zewnętrzną (weryfikacja w kontakcie z faktami).

W pojedynczym studium przypadku traktuję model jako czarną skrzynkę i stosuję trzy proste, replikowalne testy: (a) ocenę jakości argumentacji z użyciem wskaźnika jakości dyskursu (DQI), (b) kalibrację na podstawie wyniku Briera i zgodności deklarowanej pewności z trafnością, (c) samokorektę (autorewizję) po przedstawieniu dowodów przeciwnych oraz utrzymanie tej korekty w krótkim horyzoncie czasowym. Wynikiem jest profil ryzyka w ujęciu ilościowym pozwalający rozróżnić, czy problem dotyczy przede wszystkim jakości informacji, czy zdolności samoregulacji poznawczej (np. pewna, lecz błędna odpowiedź bez korekty). Zakończę omówieniem użyteczności procedury, wskazaniem jej ograniczeń oraz propozycją dalszych testów i zastosowań w badaniach medioznawczych i politologicznych.

Słowa kluczowe: sztuczna inteligencja w mediach; audyt modeli językowych; metodologia ilościowa