PISM POLSKI INSTYTUT SPRAW MIĘDZYNARODOWYCH
THE POLISH INSTITUTE OF INTERNATIONAL AFFAIRS

NR 63 (3066), 18 CZERWCA 2025 © PISM

BIULETYN

Wpływ sztucznej inteligencji na dezinformację

Aleksandra Wójtowicz

Sztuczna inteligencja jest coraz szerzej wykorzystywana przez podmioty szerzące dezinformację. Choć początkowo największe obawy wzbudzało wykorzystanie materiałów typu deepfake, obecnie rosnącym zagrożeniem staje się manipulowanie modelami generatywnymi, m.in. ingerencje w zbiory danych, na których trenuje się AI. Celem jest doprowadzenie do publikowania fałszywych treści przez modele AI. Kluczowe dla budowania odporności jest szybkie zabezpieczenie zbiorów danych, wzmocnienie moderacji oraz dostosowanie regulacji unijnych i krajowych do nowych wyzwań.

AI do tworzenia dezinformacji. Generatywna sztuczna inteligencja, czyli tworząca nowe treści, wspomaga operacje wpływu, automatyzując produkcję treści. Modele językowe są wykorzystywane do generowania spójnych narracji, fałszywych artykułów, komentarzy oraz odpowiedzi w czasie rzeczywistym w mediach społecznościowych (SM). Aktorzy dezinformacji coraz częściej używają też AI do koordynacji tzw. siatek kont trolli (autentycznych autorów) i botów (sztucznych kont zdolnych do masowej, zsynchronizowanej dystrybucji) w celu szybszego tworzenia i łatwiejszego zarządzania dezinformacyjnymi treściami w czasie rzeczywistym.

Choć początkowo największe obawy budziły materiały typu *deepfake* (wygenerowane lub zmanipulowane przez sztuczną inteligencję obrazy, treści dźwiękowe lub wideo), ich znaczenie w praktyce operacyjnej pozostaje obecnie ograniczone. Produkcja realistycznych nagrań audio czy wideo jest kosztowna i trudniejsza do rozpowszechnienia niż generowanie tekstowych narracji w SM. Badania wskazują, że w kampaniach wyborczych *deepfake* najczęściej pojawiają się w formie memów. Zamiast filmów dobrze imitujących rzeczywistość częściej wykorzystywane są materiały niskiej jakości, czyli *cheapfake*. Nie oznacza to jednak eliminacji zagrożenia ze względu na szybki rozwój technologii i wciąż wysoką skuteczność w operacjach wpływu (np. tworzeniu wiarygodnych nagrań imitujących wystąpienia kandydatów w wyborach). Technologię *deepafake* wykorzystuje się także do tworzenia tzw. person – fałszywych tożsamości dla kont trolli, które mają sprawiać wrażenie autentycznych.

Praktyczne zastosowanie tych technologii obserwujemy w rosyjskich operacjach, takich jak „Doppelganger" czy „Overload", w których AI służy do masowego tworzenia treści promujących określone narracje, często antyzachodnie lub prorosyjskie. „Doppelganger" – uznawana za jedną z największych operacji dezinformacyjnych wymierzonych przeciwko Zachodowi – polega na tworzeniu fałszywych stron internetowych i profili w SM, imitujących znane, wiarygodne media zachodnie. Publikowane na nich zmanipulowane treści mają m.in. osłabić poparcie dla Ukrainy, wzbudzić nieufność wobec instytucji Zachodu oraz wpływać na procesy wyborcze w państwach UE i NATO. „Overload" z kolei polega na masowym wysyłaniu e-maili, postów i filmów do redakcji, fact-checkerów i badaczy fałszywych lub zmanipulowanych treści, często z wykorzystaniem podrobionych logo oraz głosów ekspertów generowanych przez AI. Jej celem jest przeciążenie i dezorientacja środowisk medialnych. „Overload" podszywa się obecnie pod takie redakcje, jak Euronews, Deutsche Welle czy RMF FM.

Dezinformacja do trenowania AI. Coraz poważniejszym zagrożeniem staje się wykorzystywanie dezinformacji do celowego zanieczyszczania zbiorów danych, na których trenowane są modele AI. Ze względu na brak możliwości pełnej weryfikacji ogromnych zbiorów danych AI może się uczyć, korzystając także ze zmanipulowanych lub fałszywych treści. Już

w 2023 r. badanie Stanfordzkiego Obserwatorium Internetu wykazało, że dane treningowe AI bywają zanieczyszczone i zwierają m.in. pornografię dziecięcą. Obecnie aktorzy dezinformacji starają się ingerować w zawartość zbiorów, zarówno tworząc własne bazy danych, jak i zanieczyszczając już istniejące.

Rosja aktywnie manipuluje zbiorami, celowo wprowadzając do nich dezinformację i fałszywe narracje. Kluczowym narzędziem tej strategii jest rozbudowana sieć „Pravda", która od 2022 r. publikuje miliony artykułów na setkach stron internetowych udających lokalne portale informacyjne w dziesiątkach krajów. Treści te, często kopiowane z rosyjskich mediów państwowych i tłumaczone automatycznie, są masowo rozpowszechniane, by zwiększyć obecność prorosyjskich narracji w globalnym ekosystemie informacyjnym. Skala operacji sprawia, że generowane w jej ramach treści trafiają do otwartych zbiorów danych, z których korzystają twórcy dużych modeli językowych (LLM) i innych systemów AI. W rezultacie model przestaje rozpoznawać dezinformację – sztuczna inteligencja, napotykając tak wiele artykułów potwierdzających jakąś tezę, uznaje ją za prawdziwą i tym samym zaczyna powielać i wzmacniać rosyjską propagandę, często cytując fałszywe źródła jako wiarygodne. Badania NewsGuard z 2025 r. wykazały, że ponad 33% testowanych chatbotów powtarzało prorosyjską dezinformację, a 70% powoływało się na spreparowane artykuły.

Modele generatywne uczą się także poprzez interakcje z użytkownikami. Funkcja ta jest więc wykorzystywana przez aktorów dezinformacji do przekazywania zmanipulowanych narracji, które mimo moderacji mogą wpływać na model. Polega to na masowym przeprowadzaniu tzw. ataków *prompt injection*, które polegają na podawaniu wprowadzających w błąd informacji z wykorzystaniem promptów (zapytań i wyszukiwań w modelach generatywnej AI, np. Chat GPT), co może prowadzić do subtelnych, trudnych do wykrycia zniekształceń w generowanych treściach.

Odpowiedź na zagrożenia. W odpowiedzi na rosnące zagrożenia związane z dezinformacją i sztuczną inteligencją UE przyjęła Akt o usługach cyfrowych (DSA) oraz Akt o sztucznej inteligencji (AI Act). Choć mają one m.in. zwiększyć przejrzystość algorytmów i bezpieczeństwo użytkowników, nie w pełni uwzględniają potencjał AI do manipulacji treściami. AI Act klasyfikuje systemy AI według potencjalnego ryzyka. W przypadku modeli generatywnych, które są stworzone do interakcji z ludźmi lub tworzenia obrazów czy dźwięków, a więc takich, które mogą ułatwić podszywanie się pod innych i manipulację, wymagane jest wyraźne oznaczenie, że dane treści powstały przy użyciu AI. Natomiast w DSA sztuczna inteligencja została uznana za ryzyko systemowe, ale regulacje nie koncentrują się bezpośrednio na AI. Odpowiedź regulacyjna jest kluczowa także po to, by nadawać ton konkretnym zasadom użytkowania wprowadzanym przez firmy.

Podmioty takie jak OpenAI, Meta czy Google zmieniły zasady użytkowania, ograniczając możliwość wykorzystywania ich produktów do celów dezinformacyjnych. OpenAI zakazało korzystania z ich narzędzi w celu podszywania się pod kandydatów w wyborach i zaktualizowało polityki bezpieczeństwa, choć paradoksalnie przestało uznawać dezinformację za zagrożenie krytyczne. Meta wprowadziła w maju 2024 r. etykietowanie treści generowanych przez AI, jednak mechanizm ten opiera się głównie na dobrowolnym oznaczeniu przez użytkowników Facebooka czy Instagrama. Równocześnie tworzone są zespoły specjalistów ds. przeciwdziałania operacjom wpływu oraz rozbudowywane mechanizmy moderacji treści.

Wnioski i rekomendacje. Sztuczna inteligencja nie tylko zmienia sposób prowadzenia operacji dezinformacyjnych, ale sama staje się ich celem. Współczesne modele językowe, mimo ogromnego potencjału, narażone są na manipulacje, co może osłabiać ich wiarygodność i pogłębiać chaos informacyjny. Odpowiedź na te wyzwania musi być kompleksowa, łącząc zmiany regulacyjne, działania technologiczne, edukację użytkowników oraz ścisłą współpracę międzysektorową. Nadal brakuje jednak skutecznej współpracy z platformami i mechanizmów szybkiego reagowania na incydenty dezinformacyjne z udziałem AI. Co więcej, niedawna zmiana podejścia platform społecznościowych do moderacji treści w imię ochrony szeroko rozumianej „wolności słowa" utrudnia walkę z dezinformacją z udziałem AI.

Dla UE kluczowe jest dostosowanie DSA do zagrożeń wynikających z wpływu aktorów dezinformacji na modele generatywne. Potrzebne jest jasne określenie obowiązków platform cyfrowych wobec treści dezinformacyjnych tworzonych przy pomocy AI. Choć OpenAI jako bardzo duża platforma internetowa (VLOP) już podlega przepisom DSA, konieczne jest doprecyzowanie w regulacji obowiązków moderacji treści w modelach AI.

Równie istotna jest intensyfikacja dialogu między platformami cyfrowymi, państwami i instytucjami UE w celu wypracowania szybkich ścieżek reagowania na dezinformację związaną z AI oraz prowadzenia skoordynowanych operacji ograniczających jej rozprzestrzenianie się. Przykładem takich działań są powoływane przez platformy centra operacyjne ds. wyborów, współpraca z fact-checkerami oraz wdrażanie mechanizmów oznaczania i usuwania treści generowanych przez AI, które mogą wprowadzać w błąd użytkowników lub zagrażać procesom demokratycznym. Dotychczas były one prowadzone przez platformy społecznościowe. W większym stopniu należy wdrożyć ich stosowanie również przez firmy rozwijające AI, w kontekście modeli generatywnych.

Wzmocnienia wymaga kontrola nad danymi treningowymi modeli AI i korzystanie z podwójnego procesu weryfikacji baz danych. Powinna ona obejmować audyty jakości i pochodzenia danych wejściowych prowadzone przez firmy zewnętrzne na zlecenie platform, jak również wdrażanie narzędzi AI trenowanych na węższych i wyspecjalizowanych zbiorach danych, zdolnych do automatycznego wykrywania fałszywych lub zmanipulowanych treści w zbiorach treningowych. Istotne, by taka kontrola była wymagana prawem UE na zasadach podobnych jak moderacja treści w DSA. Platformy musiałyby więc wytworzyć odpowiednie mechanizmy, a w przypadku ich niewdrożenia mogłyby zostać pozwane przez Komisję Europejską.

Redakcja: Jarosław Ćwiek-Karpowicz, Rafał Tarnogórski, Marcin Terlikowski, Justyna Szczudlik
Jędrzej Czerep, Amanda Dziubińska, Wojciech Lorenz, Mateusz Piotrowski, Daniel Szeligowski, Jolanta Szymańska, Damian Wnukowski, Szymon Zaręba, Tomasz Żornaczuk