Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji

dla mediów i twórców cyfrowych

NASK

Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji

dla mediów i twórców cyfrowych

Sylwia Adamczyk (red.)

Ewelina Bartuzi-Trokielewicz

Bartłomiej Dobosiewicz

Agnieszka Karlińska

Anna Kołos

Kornel Kowieski

Anna Kwaśnik

Patrycja Krzyśpiak

Michał Marek

Alicja Martinek

Konrad Marzęcki

Karolina Seweryn

Olga Wałkuska

NASK

Ministerstwo Cyfryzacji

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

Tytuł
Przewodnik po cyberbezpieczeństwie i sztucznej inteligencji
dla mediów i twórców cyfrowych

Autorzy
Sylwia Adamczyk (red.), Ewelina Bartuzi-Trokielewicz, Bartłomiej Dobosiewicz, Agnieszka Karlińska, Anna Kołos, Kornel Kowieski, Anna Kwaśnik, Patrycja Krzyśpiak, Michał Marek, Alicja Martinek, Konrad Marzęcki, Karolina Seweryn, Olga Wałkuska

Wsparcie
Zuzanna Polak

Redakcja językowa
Katarzyna Nakonieczna, Łukasz Szczęsny

Opracowanie graficzne
NASK – PIB

ISBN: 978-83-68356-20-5

2025

NASK – Państwowy Instytut Badawczy
ul. Kolska 12, 01–045 Warszawa,
www.nask.pl

Spis treści

Wstęp

Współcześni dziennikarze funkcjonują w środowisku informacyjnym, które zmienia się szybciej niż kiedykolwiek wcześniej. Rozwój technologii cyfrowych, sztucznej inteligencji oraz mediów społecznościowych otworzył nowe możliwości w zakresie pozyskiwania, tworzenia i dystrybucji informacji. Jednocześnie jednak przyniósł ze sobą szereg wyzwań i zagrożeń, które wymagają od branży dziennikarskiej nie tylko wysokich kompetencji merytorycznych i umiejętności wykorzystania potencjału nowoczesnych technologii, ale również świadomości z zakresu cyberbezpieczeństwa i zdolności do rozpoznawania pułapek związanych z tymi narzędziami.

Celem niniejszej publikacji jest przedstawienie kluczowych zagadnień związanych z cyberbezpieczeństwem w pracy dziennikarza/dziennikarki, a także omówienie wpływu nowoczesnych technologii – takich jak duże modele językowe (LLM), czy deepfejki – na jakość i wiarygodność informacji. Wskazujemy również na konkretne zagrożenia, takie jak dezinformacja, manipulacja obrazem i dźwiękiem, czy działania aktorów państwowych w infosferze, ze szczególnym uwzględnieniem rosyjskiej i białoruskiej propagandy w polskojęzycznym internecie. Warto również podkreślić prawne aspekty wykorzystania AI w pracy dziennikarza/dziennikarki w kontekście odpowiedzialności czy praw autorskich, co w przyszłości będzie regulował Akt o Sztucznej Inteligencji.

Publikacja ma charakter praktyczny – zawiera nie tylko analizę zjawisk, ale także zestaw rekomendacji i dobrych praktyk, które mogą pomóc dziennikarzom i twórcom treści w bezpiecznym i odpowiedzialnym korzystaniu z narzędzi cyfrowych. Mamy nadzieję, że przyczyni się do utrzymania wysokiej jakości przekazu medialnego.

ZUZANNA POLAK

ANNA KWAŚNIK

Cyberhigiena w pracy dziennikarzy

01.

Współcześni dziennikarze działają na przecięciu technologii, informacji i zaufania społecznego, co czyni media – zarówno ogólnopolskie, jak i lokalne – atrakcyjnym celem dla cyberprzestępców. Redakcje mają realny wpływ na opinię publiczną i przechowują cenne dane, a przejęcie ich kanałów komunikacji może być wykorzystane do dezinformacji i destabilizacji. Zagrożone są nie tylko duże redakcje – mniejsze, lokalne media, ze względu na ograniczone zasoby, często stają się łatwym celem ataków i narzędziem wpływu na konkretne społeczności.

Dane gromadzone przez redakcje i dziennikarzy – potencjalny cel cyberataków

Dziennikarze i redakcje przetwarzają szeroki zakres danych, które z punktu widzenia cyberprzestępców mają nie tylko wysoką wartość informacyjną, ale też strategiczną i operacyjną. Wśród nich znajdują się zarówno dane osobowe, materiały redakcyjne, jak i informacje pozyskane w toku pracy śledczej.

Do najczęściej przetwarzanych i potencjalnie atrakcyjnych, a tym samym zagrożonych danych należą:

- Dane osobowe i kontaktowe rozmówców, ekspertów i współpracowników, takie jak adresy e-mail, numery telefonów, adresy zamieszkania, stanowiska czy treści korespondencji służbowej. Dane te mogą być użyte do kradzieży tożsamości, ataków phishingowych, socjotechnicznych, a także do dalszych działań wymierzonych w konkretne osoby lub instytucje.

- Dane informatorów, które są objęte tajemnicą dziennikarską i przekazywane w zaufaniu. W tej kategorii mieszczą się m.in. imiona i nazwiska (często (za)szyfrowane), dane kontaktowe, nagrania rozmów, zapisy czatów oraz dokumentacja dowodowa. Ujawnienie tych danych może narazić informatorów na represje zawodowe, społeczne, a w skrajnych przypadkach – na realne zagrożenie życia lub zdrowia.
- Nieopublikowane materiały dziennikarskie takie jak notatki, zdjęcia, artykuły, nagrania audio i wideo, dane statystyczne czy analizy. Często stanowią one efekt wielomiesięcznej pracy dziennikarzy, zwłaszcza śledczych, i są podstawą publikacji o dużym znaczeniu społecznym lub politycznym. W wielu przypadkach są też objęte embargiem lub tajemnicą redakcyjną. Ich przejęcie może posłużyć manipulacji przekazem, szerzeniu dezinformacji, zablokowaniu publikacji lub szantażowi wobec dziennikarzy i redakcji.
- Dane wrażliwe dotyczące instytucji publicznych i prywatnych, takie jak sprawozdania finansowe, raporty z audytów, materiały ujawniające nieprawidłowości, nadużycia czy korupcję. Tego typu dane, zdobywane w toku dziennikarstwa śledczego, mają znaczenie polityczne, operacyjne lub gospodarcze i mogą być wykorzystane do skompromitowania instytucji, zastraszania osób zaangażowanych lub uzyskania przewagi konkurencyjnej.

Ze względu na zakres i charakter przechowywanych informacji, redakcje stanowią atrakcyjny cel dla cyberprzestępców – zarówno tych działających dla zysku finansowego, jak i podmiotów zaangażowanych w działania wywiadowcze, sabotaż informacyjny czy szerzenie dezinformacji. Dlatego bezpieczeństwo danych powinno być w mediach traktowane z najwyższym priorytetem – na równi z etyką zawodową i rzetelnością dziennikarską.

Infrastruktura cyfrowa i wizerunek redakcji jako obiekty ataków

Funkcjonowanie współczesnych redakcji opiera się na rozbudowanej infrastrukturze informatycznej oraz cyfrowych kanałach dystrybucji opracowywanych przez nie treści. Potencjalny atak na te elementy może sparaliżować działalność redakcji, naruszyć zaufanie odbiorców, a przede wszystkim posłużyć do manipulowania przekazem informacyjnym.

1. Infrastruktura IT jako cel ataku

Redakcje korzystają z wielu narzędzi i systemów wspierających pracę zespołu – m.in. serwerów, baz danych, systemów CMS, platform do przesyłania materiałów oraz poczty elektronicznej – które stanowią krytyczne elementy funkcjonowania każdej organizacji medialnej.

Zagrożenia związane z naruszeniem infrastruktury IT obejmują:

- zahamowanie procesów redakcyjnych,
- zaszyfrowanie lub utratę danych (np. w wyniku ataku ransomware),
- wyciek niepublikowanych treści i informacji,
- przejęcie poufnej korespondencji z partnerami, źródłami lub innymi redakcjami.

2. **Kanały dystrybucji treści atrakcyjnym celem dla cyberprzestępców**
Strony internetowe, konta w mediach społecznościowych, podcasty, newslettery i platformy wideo to podstawowe narzędzia dotarcia do odbiorców. To właśnie one stanowią główny kanał komunikacji z opinią publiczną.

Przejęcie kanałów dystrybucji przez osoby nieuprawnione może skutkować:

- publikacją fałszywych lub zmanipulowanych treści,
- kompromitacją redakcji i utratą wiarygodności,
- wywołaniem chaosu informacyjnego pod marką zaufanego medium,
- podszywaniem się pod dziennikarzy lub zespół redakcyjny w celu rozprzestrzeniania dezinformacji.

3. **Społeczne zaufanie jako zasób strategiczny**
Najważniejszym kapitałem mediów jest zaufanie społeczne – zdobywane latami, budowane poprzez rzetelność, profesjonalizm i niezależność. To zaufanie umożliwia wpływ na debatę publiczną i legitymizuje działalność informacyjną.

Dlatego celem cyberataków bywa nie tylko fizyczny dostęp do danych, lecz także:

- osłabienie reputacji redakcji
- zakłócenie relacji z odbiorcami
- wykorzystanie marki medialnej do realizacji celów politycznych, ideologicznych lub ekonomicznych
- tworzenie i rozpowszechnianie dezinformacji pod szyldem wiarygodnego źródła.

Naruszenie infrastruktury cyfrowej redakcji i zaufania społecznego może mieć skutki znacznie poważniejsze niż krótkotrwała utrata danych – może podważyć fundament niezależności medialnej i zachwiać jej rolą w porządku informacyjnym.

Atak na Polską Agencję Prasową

31 maja 2024 roku Polska Agencja Prasowa (PAP) padła ofiarą poważnego cyberataku[1]. W jej serwisie informacyjnym opublikowano dwie fałszywe depesze, informujące o rzekomej decyzji premiera Donalda Tuska o częściowej mobilizacji 200 tysięcy obywateli i wysłaniu ich do Ukrainy. Mimo że treści te zostały szybko usunięte z serwisu PAP[2], zostały podchwycone przez inne media, co stworzyło ryzyko eskalacji dezinformacji i wywołania społecznej paniki.

Jak ujawniono, atak był efektem wcześniejszego uzyskania przez cyberprzestępców dostępu do konta jednego z pracowników PAP. Prawdopodobnie dokonano tego poprzez zainfekowanie jego komputera złośliwym oprogramowaniem, co pozwoliło na przejęcie danych uwierzytelniających i publikację nieautoryzowanych treści.

Reakcja PAP była szybka, odpowiedzialna i transparentna – natychmiast po wykryciu incydentu agencja wydała komunikat, w którym jasno przekazała, że nie była autorem informacji. Jednocześnie zidentyfikowano i zablokowano drogę, którą posłużono się do przeprowadzenia ataku. W działania zaangażowano służby państwowe, w tym Agencję Bezpieczeństwa Wewnętrznego i Ministerstwo Cyfryzacji. Prokuratura Okręgowa w Warszawie wszczęła śledztwo w sprawie prób dezinformacji i destabilizacji państwa.

Ten przypadek pokazuje, jak ważne jest posiadanie mechanizmów szybkiego reagowania i komunikacji kryzysowej. PAP, mimo skali zagrożenia, zareagowała sprawnie: zabezpieczyła system, poinformowała odbiorców i przekazała sprawę odpowiednim organom. Takie działanie jest modelowym przykładem tego, jak media powinny postępować w przypadku naruszenia bezpieczeństwa cyfrowego.

Bezpieczeństwo jako fundament niezależnych mediów

Redakcje i dziennikarze operują danymi o wysokim stopniu wrażliwości – od informacji osobowych i materiałów dziennikarskich, po dane o źródłach i instytucjach. Ich przejęcie może prowadzić do naruszenia prywatności, manipulacji treściami i utraty kontroli nad przekazem. Równie ważne jest zabezpieczenie infrastruktury IT i kanałów komunikacji – naruszenie ich integralności grozi paraliżem pracy i utratą zaufania odbiorców. Ochrona danych i systemów to dziś klucz do bezpieczeństwa informacyjnego i zachowania niezależności redakcji.

Dlaczego cyberhigiena jest niezbędna w pracy dziennikarskiej?

W obliczu coraz częstszych incydentów związanych z cyberbezpieczeństwem – również w środowisku medialnym – zasady cyberhigieny nie mogą być traktowane jako techniczny dodatek. Ich stosowanie staje się nieodzownym elementem codziennej pracy redakcji. To warunek ochrony danych, zapewnienia ciągłości działań oraz utrzymania zaufania odbiorców. Cyberhigiena stanowi integralną część warsztatu każdego odpowiedzialnego dziennikarza/dziennikarki.

Praktyczne wskazówki i mechanizmy ochrony

1. Bezpieczne hasła i uwierzytelnianie

- Każdy pracownik redakcji powinien korzystać z silnych, unikalnych haseł.
- Wszędzie tam, gdzie to możliwe, należy stosować uwierzytelnianie dwuskładnikowe (2FA), zwłaszcza do systemów CMS, poczty redakcyjnej i kont w mediach społecznościowych.

2. Zarządzanie sprzętem i oprogramowaniem

- Oprogramowanie wszystkich urządzeń wykorzystywanych do pracy dziennikarskiej powinno być regularnie aktualizowane – zarówno system operacyjny, jak i poszczególne aplikacje, w szczególności programy antywirusowe.
- Nośniki danych (pendrive'y, dyski zewnętrzne) powinny być szyfrowane i stosowane z rozwagą. Warto wprowadzić politykę, która jasno określa zasady używania nośników zewnętrznych.
- Urządzenia mobilne oraz komputery używane do celów służbowych powinny być zabezpieczone hasłem, kodem PIN lub biometrią.

3. Kopie zapasowe i dostęp do danych

- Wszystkie kluczowe dane – zarówno materiały dziennikarskie, jak i dane źródeł – powinny być regularnie archiwizowane w bezpieczny sposób, najlepiej w chmurze oraz na nośnikach offline. Zaleca się stosowanie strategii 3-2-1 (3 kopie danych, 2 różne nośniki, 1 kopia w innej lokalizacji) oraz regularne testowanie odzyskiwania danych z kopii zapasowych[3].
- Warto stosować szyfrowanie plików i dysków, zwłaszcza w przypadku materiałów wrażliwych lub objętych embargiem.
- Należy wdrożyć zasadę ograniczonego dostępu (tzw. zasada najmniejszych uprawnień): pracownicy redakcji powinni mieć dostęp wyłącznie do tych danych

i systemów, które są im niezbędne do wykonywania obowiązków służbowych. Dzięki temu minimalizuje się ryzyko nieuprawnionego wykorzystania informacji oraz ogranicza skutki ewentualnego naruszenia.

4. Oddzielanie spraw służbowych od prywatnych

Jednym z podstawowych i często niedocenianych aspektów cyberhigieny w pracy dziennikarskiej jest wyraźne oddzielenie przestrzeni zawodowej od prywatnej – zarówno w zakresie sprzętu, jak i kont oraz profili w mediach społecznościowych. Łączenie tych sfer zwiększa ryzyko wycieku danych, przypadkowego ujawnienia informacji czy wykorzystania służbowych zasobów do celów nieautoryzowanych.

REKOMENDOWANE PRAKTYKI:

- Urządzenia – sprzęt redakcyjny (m.in. laptopy, telefony, nośniki danych) powinien być używany wyłącznie do celów zawodowych. Unikanie logowania się na prywatne konta (np. poczta, bankowość, media społecznościowe) na urządzeniach służbowych znacząco ogranicza ryzyko przeniesienia zagrożeń z jednej sfery do drugiej.
- Konta e-mail – dziennikarze powinni korzystać z oddzielnych adresów do komunikacji zawodowej. Wszelka korespondencja redakcyjna, kontakt z informatorami, przesyłanie materiałów i logowanie do systemów powinno odbywać się wyłącznie z kont firmowych lub kont zabezpieczonych na potrzeby pracy dziennikarskiej.
- Profile w mediach społecznościowych – nie należy zarządzać kontami redakcyjnymi z prywatnych urządzeń ani mieszać ról administratora redakcji z kontami osobistymi. Wskazane jest stosowanie narzędzi typu menedżer stron, z jasno przydzielonymi uprawnieniami oraz ograniczenie dostępu wyłącznie do osób odpowiedzialnych za komunikację.
- Chmura i przechowywanie plików – materiały redakcyjne nie powinny być przechowywane na prywatnych kontach w usługach chmurowych, o ile nie są one objęte dodatkowymi warstwami zabezpieczeń i formalną kontrolą dostępu.

5. Bezpieczna komunikacja

- Komunikacja – zarówno wewnętrzna, jak i zewnętrzna – powinna odbywać się wyłącznie za pośrednictwem szyfrowanych kanałów, takich jak komunikatory z szyfrowaniem end-to-end lub zaszyfrowane wiadomości e-mail (np. z użyciem protokołu PGP).
- Dziennikarze powinni unikać przesyłania wrażliwych informacji przez popularne, niezabezpieczone platformy (np. standardowy SMS, prywatny e-mail, komunikatory).

- Należy zachować ostrożność przy otwieraniu załączników i linków w e-mailach – phishing pozostaje jednym z najczęstszych wektorów ataku.

6. Ochrona przed phishingiem

W przypadku wiadomości e-mail, SMS-ów oraz komunikatorów należy zachować szczególną ostrożność – zwłaszcza jeśli zawierają linki lub załączniki, a ich treść wzbudza emocje lub wymaga podjęcia natychmiastowych działań. Nawet gdy treść wydaje się wiarygodna, należy stosować podstawowe zasady bezpieczeństwa:

- Weryfikować pełny adres nadawcy, a nie tylko nazwę wyświetlaną w polu „od". Fałszywe adresy często różnią się jedną literą lub pochodzą z nietypowych domen.
- Unikać klikania w linki bez uprzedniego sprawdzenia ich wiarygodności, szczególnie gdy wiadomość wywołuje presję czasu, zawiera prośbę o dane logowania lub żąda natychmiastowej reakcji.
- Zwracać uwagę na załączniki w wiadomościach od nieznanych nadawców, szczególnie na rozszerzenie pliku. Przestępcy często wykorzystują pliki typu .tar, .exe, .js czy .7z do dystrybucji złośliwego oprogramowania.

W przypadku wątpliwości zaleca się potwierdzenie autentyczności wiadomości innym, znanym kanałem komunikacji. Zachowanie czujności i ograniczonego zaufania stanowi skuteczną barierę przed atakami phishingowymi.

7. Edukacja i świadomość zespołu

- Szkolenia dla pracowników z zakresu cyberbezpieczeństwa – nie tylko technicznego, ale również praktycznego (np. rozpoznawanie prób socjotechniki) – należy realizować cyklicznie, a nowi pracownicy powinni przechodzić wstępne szkolenie bezpieczeństwa jako część wdrożenia do pracy.
- Warto wprowadzić jasne procedury reagowania na incydenty – kto zgłasza, komu i w jakim trybie.

8. Procedury bezpieczeństwa i reagowania

- Należy wprowadzić politykę bezpieczeństwa informacji, która zawiera wytyczne dotyczące przechowywania, przesyłania i zabezpieczania danych.
- Redakcja powinna posiadać przygotowany plan reagowania na incydenty, który obejmuje identyfikację i zgłoszenie naruszenia, prowadzenie komunikacji kryzysowej (zarówno wewnętrznej, jak i zewnętrznej), analizę ryzyka i usuwanie skutków zdarzenia oraz zgłoszenie incydentu do odpowiednich organów.

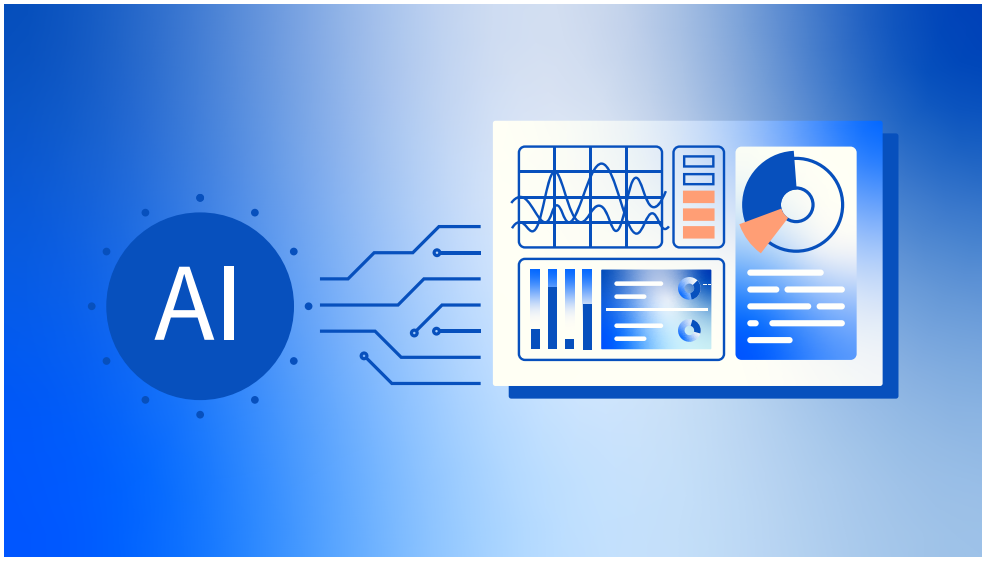
Współczesne redakcje i dziennikarze, jako podmioty szczególnie narażone na cyberataki, muszą działać z najwyższą ostrożnością i świadomością zagrożeń. W dobie powszechnych prób dezinformacji i ingerencji w przekaz medialny nie mogą pozwolić sobie na pośpiech, rutynę ani lekceważenie podstawowych zasad bezpieczeństwa. Wdrożenie cyberhigieny nie wymaga specjalistycznej wiedzy technicznej, lecz konsekwencji, uważności i stosowania sprawdzonych praktyk. To nie tylko sposób na ochronę danych, ale fundament wiarygodności, niezależności i stabilności pracy dziennikarskiej.

ANNA KWAŚNIK

KAROLINA SEWERYN, ANNA KOŁOS,
AGNIESZKA KARLIŃSKA

LLM-y w generowaniu treści: ryzyka i zasady bezpiecznej pracy

02.



Duże modele językowe (ang. Large Language Models, w skrócie LLM) to systemy, które potrafią generować tekst w języku naturalnym i rozwiązywać różnorodne zadania związane z przetwarzaniem tekstu, takie jak streszczanie, parafrazowanie, tłumaczenie czy tworzenie wypowiedzi w określonym gatunku. Ich działanie polega na przewidywaniu kolejnych słów w zdaniu na podstawie wzorców zaobserwowanych podczas treningu na ogromnych zbiorach danych tekstowych. Do najbardziej znanych modeli należą m.in. ChatGPT, Claude, Gemini, Mistral czy LLaMa. Coraz większą rolę odgrywają też inicjatywy lokalne, które dążą do tworzenia modeli dostosowanych do specyfiki językowej, kulturowej i historycznej poszczególnych krajów. W Polsce takimi projektami są PLLuM i Bielik, których celem jest rozwój otwartych modeli językowych lepiej odzwierciedlających polski kontekst. Szerokie możliwości LLM-ów są już całkiem dobrze znane w świecie mediów, dlatego w tekście skupimy się na omówieniu ich ograniczeń.

Halucynacje

Duże modele językowe mają skłonność do tzw. halucynacji, czyli generowania odpowiedzi, które brzmią wiarygodnie i przekonująco, ale są nieprawdziwe[4,5]. To zjawisko jest nieodłączną cechą działania LLM-ów, które nie umieją odróżnić prawdy od fikcji. W efekcie potrafią „wymyślić” cytaty, dane liczbowe, instytucje, wydarzenia, a nawet osoby czy publikacje, które nigdy nie istniały. Halucynacje mogą przybierać różne formy – od drobnych nieścisłości po całkowicie zmyślone treści.

Problem pogłębia fakt, że modele zazwyczaj nie przyznają się do braku wiedzy. Wręcz przeciwnie – są zaprogramowane, by zawsze udzielić odpowiedzi, nawet jeśli powinny odmówić. Co więcej, ludzie oceniający odpowiedzi modeli językowych często preferują konkrety i pewność siebie modelu, więc algorytm uczy się unikania oznak niepewności. Trwają prace nad technikami, które zachęcą modele do przyznawania się do niewiedzy (np. *R-Tuning*[6]), jednak to wciąż wyzwanie badawcze.

Halucynacje LLM stanowią poważne zagrożenie w kontekście dziennikarstwa i publikacji medialnych. Model potrafi tworzyć fikcyjne informacje i opisywać wydarzenia, które nigdy nie miały miejsca, co może prowadzić do szerzenia fake newsów. Szczególnie niebezpieczną odmianą halucynacji są sytuacje, gdy model fabrykuje przypisy, cytaty lub źródła, które wyglądają wiarygodnie, lecz nie istnieją. Taka „halucynacja źródłowa" może zwieść użytkownika, że informacja jest poparta dowodem. Mylące lub fałszywe treści wygenerowane przez modele sztucznej inteligencji, jeśli trafiłyby na łamy gazet czy portali, mogłyby nadszarpnąć reputację redakcji. Dziennikarze korzystający z pomocy LLM muszą zachować najwyższą ostrożność i każdą informację wygenerowaną przez system sztucznej inteligencji traktować jak niezweryfikowane źródło.

W ciągu ostatnich dwóch lat odnotowano wiele pomyłek generowanych przez AI, które unaoczniły skalę problemu. Oto dwa przykłady z różnych dziedzin:

- **Fałszywe oskarżenia o morderstwo.** W Norwegii ChatGPT wygenerował szczegółową, rzekomo „prawdziwą" historię o Arve Holmenie – opisując go jako ojca, który zamordował dwoje swoich dzieci i został za to skazany na 21 lat więzienia[7]. W rzeczywistości Holmen nigdy nie popełnił żadnej z tych zbrodni – historia była całkowicie zmyślona.
- **Fałszywe cytowania prawne.** W Stanach Zjednoczonych Ameryki kancelaria prawna została ukarana grzywną za oparcie się na nieistniejących cytatach i danych procesowych wygenerowanych przez AI, co doprowadziło do wprowadzenia w błąd sądu[8].

Dezinformacja

Ze zjawiskiem halucynacji powiązane są dwa istotne ryzyka: ryzyko wykorzystania LLM-ów – tym razem już w sposób w pełni intencjonalny – do rozpowszechniania dezinformacji oraz ryzyko manipulacji danymi treningowymi w celu kształtowania narracji propagandowych.

Dezinformacja z wykorzystaniem LLM-ów

Duże modele językowe są mieczem obosiecznym – mogą być skutecznym narzędziem zarówno w walce z dezinformacją (na bazie LLM-ów powstaje coraz więcej rozwiązań do wykrywania fake newsów), jak i w jej tworzeniu[9,10,11]. Jak już zostało powiedziane, LLM-y potrafią generować teksty brzmiące bardzo naturalnie, a ich forma i styl mogą wywoływać wrażenie wiarygodności. Chociaż popularne modele komercyjne mają wbudowane zabezpieczenia, które istotnie ograniczają możliwości nielegalnego użycia, ryzyka tworzenia za ich pomocą fake newsów nie da się w pełni wyeliminować. Z kolei modele otwartożródłowe można prościej dostosowywać do konkretnych potrzeb, ale niestety łatwiej dostosować je również do generowania treści nielegalnych. Jak wskazują badania, teksty generowane przez LLM-y są coraz szerzej wykorzystywane w kampaniach dezinformacyjnych i propagandowych w sposób celowany, np. w określonych językach i na konkretnych platformach społecznościowych[12]. Technologie SI umożliwiają znacznie efektywniejsze generowanie fake newsów i pokrycie większej liczby tematów bez obniżania siły przekonywania ani wiarygodności komunikatów w oczach odbiorców[13]. Wykazano, że fałszywe treści tworzone przez LLM-y są trudniejsze do wykrycia niż te tworzone przez ludzi, a ich wiarygodność wynika z kontekstowego dopasowania oraz wysokiej spójności językowej[14].

Manipulacja danymi treningowymi i zjawisko LLM grooming

Manipulacja danymi treningowymi, w szczególności celowe wprowadzanie treści, które mają wpłynąć na wyniki generowane przez modele – zjawisko określane często jako LLM grooming – to kolejny, mniej oczywisty i niestety jeszcze trudniejszy do wykrycia sposób na wykorzystanie LLM-ów do promowania określonych narracji. Treści generowane przez zmodyfikowane w ten sposób modele językowe mogą wydawać się wiarygodne nawet dla wyspecjalizowanych narzędzi do wykrywania dezinformacji. Jako przykład wskazać można operację moskiewskiej sieci dezinformacyjnej „Pravda", której celem było manipulowanie treściami generowanymi przez zachodnie modele językowe w kierunku narracji prorosyjskich i antyzachodnich. Sieć stosowała techniki SEO, aby podnieść widoczność zmanipulowanych materiałów w wynikach wyszukiwania. W efekcie popularne czatboty SI częściej korzystały właśnie z tych danych, powielając fałszywe informacje o prorosyjskim wydźwięku i niekiedy nawet cytując propagandowe artykuły jako źródła[15,16,17,18].

Opisane zjawiska czynią z modeli językowych niebezpieczne narzędzia w rękach rządów autorytarnych czy grup przestępczych. Przedstawiciele i przedstawicielki mediów powinni być szczególnie wyczuleni na treści generowane przez LLM-y dotyczące kluczowych tematów politycznych i społecznych, które pozornie wydają się spójne i wiarygodne, a w rzeczywistości mogą być elementem skoordynowanych kampanii manipulacyjnych.

Stronniczość

Korzystając z modeli językowych, trzeba mieć świadomość, że mogą w nich być zaszyte obciążenia (ang. biases) czy preferencje światopoglądowe[19,20]. Obecność stereotypów i uprzedzeń wynika przede wszystkim z tego, na jakich danych model był uczony. LLM-y odzwierciedlają strukturę danych treningowych – jeśli dane te zawierają treści seksistowskie, rasistowskie, homofobiczne lub inne stanowiące wyraz uprzedzeń wobec określonych grup społecznych, istnieje duże ryzyko, że model będzie powielał te wzorce w swoich odpowiedziach. Jednym z dobrze udokumentowanych przykładów takich uprzedzeń jest gender bias w zadaniu tłumaczenia maszynowego: modele często wybierają formę męską jako domyślną, ignorując kontekst, który mógłby wskazywać na użycie formy żeńskiej. Innym, nowszym przykładem są stronnicze odpowiedzi chińskich modeli Qwen na pytania o wydarzenia z Placu Tian'anmen[21].

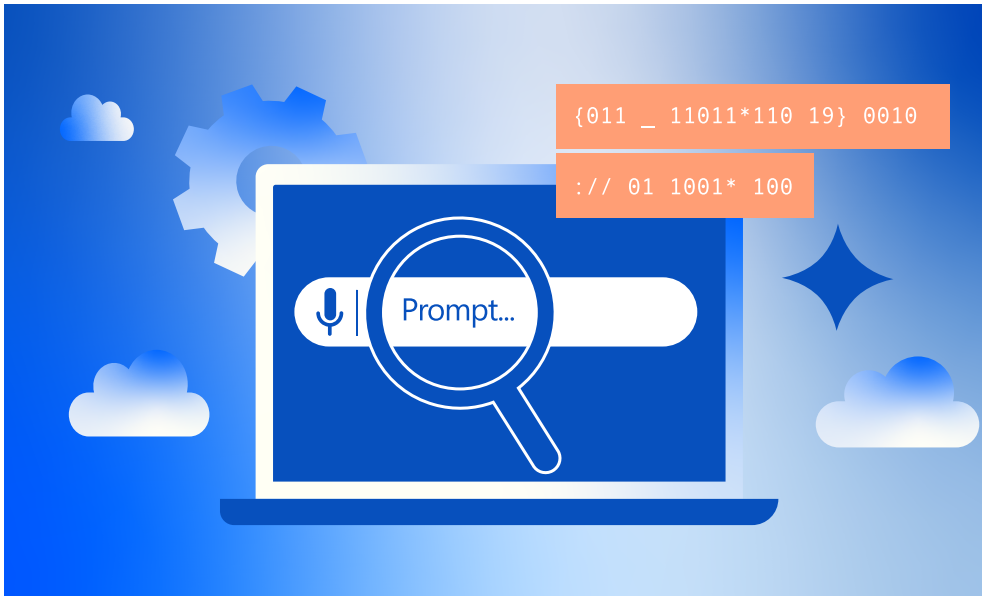
Problemem jest nie tylko sam fakt reprodukcji uprzedzeń, ale również to, że modele mogą wzmacniać je w sposób trudny do przewidzenia, np. przez tworzenie treści jeszcze bardziej skrajnych niż te obecne w danych źródłowych. Zjawisko to jest szczególnie niebezpieczne, gdy modele są wykorzystywane w sytuacjach, które wymagają obiektywności i równego traktowania, np. w administracji publicznej czy w procesach rekrutacyjnych.

W wielu LLM-ach, np. w modelach z rodziny PLLuM, stosuje się techniki redukcji uprzedzeń, takie jak wychowanie (ang. alignment), a więc dostosowanie modeli do wartości i norm społecznych oraz ich zabezpieczenie przed generowaniem treści szkodliwych. Wychowanie jest jednak procesem, który ma swoje ograniczenia – modele nie tylko „uczą się" określonych wartości, ale także mogą przejawiać skrzywienia wynikające z ich specyficznej interpretacji. W efekcie przykładowy model może rzadziej klasyfikować jako obraźliwe treści zgodne z przekonaniami wpojonymi na etapie wychowania. Z tego powodu konieczne jest krytyczne analizowanie treści generowanych przez modele językowe pod kątem nie tylko wiarygodności informacji, lecz także wyważenia i obecności obciążeń, tych mniej i bardziej subtelnych.

Język

Mimo imponujących zdolności dużych modeli językowych do generowania tekstu, również z uwzględnieniem specyficznych wytycznych odnośnie do formatowania i określonych konwencji, musimy zdawać sobie sprawę z ograniczeń w zakresie zarówno bogactwa i zróżnicowania językowego, jak i poprawności sensu stricto, zwłaszcza kiedy mowa o językach słabiej reprezentowanych w danych treningowych sztucznej inteligencji. Najbardziej konkurencyjne modele na rynku globalnym trenowane są na zbiorach, w których 80–90% stanowią dane anglojęzyczne, a udział mniejszych języków z reguły nie przekracza kilku procent. Nawet jednak w odniesieniu do powszechnie dominującego języka angielskiego warto pamiętać, że model został nauczony naśladowania konwencji i wzorców reprezentowanych w ogromnych zbiorach treningowych, jego domeną nie jest natomiast oryginalność czy kreatywność, polegające na twórczych odstępstwach od norm. Od dwóch lat w internecie krążą kompilacje fraz i pojedynczych słów nadużywanych przez modele typu GPT, które stanowią cechę charakterystyczną tekstów niestworzonych przez człowieka.

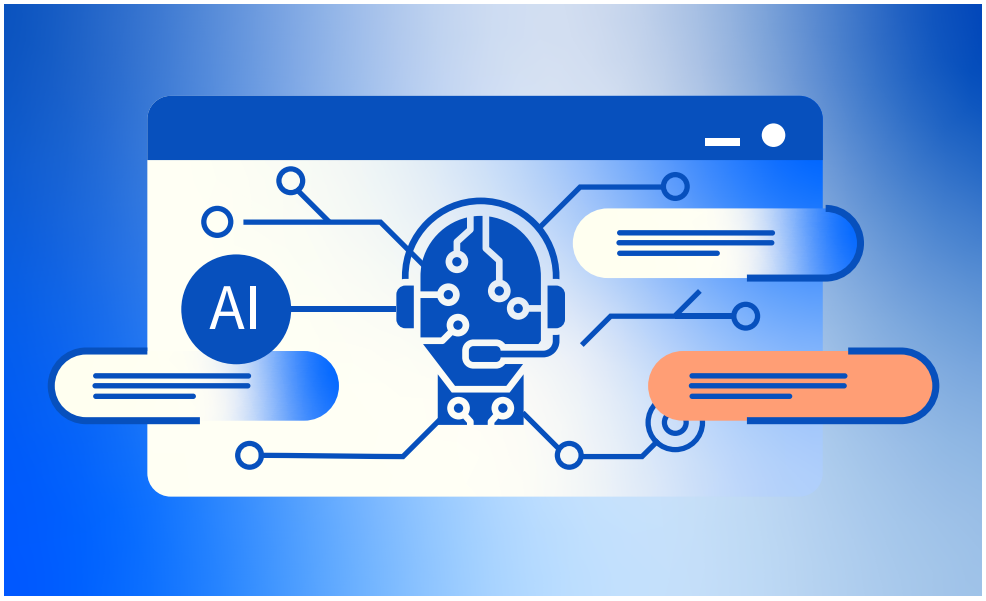
Angielskie leksemy, takie jak rzeczownik *tapestry* czy czasownik *delve*, stały się emblematycznymi przykładami obserwowanych dysproporcji między częstotliwością użycia w odpowiedziach ChatGPT a ich niską frekwencją w tekstach pisanych przez ludzi. „The Guardian" twierdzi nawet, że przyczyny nadużywania takich słów jak *delve* leżą w nierównościach ekonomicznych leżących u podstaw trenowania modeli. Czasownik *delve* ma być charakterystyczny dla angielskiego języka biznesowego w Nigerii, wskazując na szerokie wykorzystanie nisko opłacanej pracy anotatorów (osób odpowiedzialnych za ręczne przygotowanie danych służących do dostrajania modeli, np.



ocenianie poprawności odpowiedzi generowanych przez model, tworzenie pytań i wzorcowych odpowiedzi) w Afryce[22]. Jeremy Nguyen ze Swinburne University of Technology w Melbourne przeprowadził analizę akademickich artykułów z zakresu medycyny, wskazując, że częstotliwość użycia słowa delve w 2024 roku wzrosła nawet stukrotnie względem lat poprzedzających upublicznienie ChatGPT. Rozwiązanie problemu powtarzalnych i budzących pewną konsternację środków językowych nadużywanych przez duże modele językowe jawi się dość paradoksalnie. Na rynku pojawiły się już bowiem cyfrowe narzędzia „humanizujące" teksty stworzone przez modele.

Problemy językowe namnażają się oczywiście, kiedy spojrzymy na zdolności modeli do tworzenia tekstu w języku polskim. Co prawda następuje na tym polu ogromny postęp, bo jeszcze dwa lata temu GPT-3.5 twierdził uparcie, że liczba mnoga od słowa dodatek to dodateki i nie pozwalał się wyprowadzić z błędu. Jednak, mimo że te czasy są już za nami, a błędy fleksyjne dotyczą zaledwie niewielkiego odsetka słów rzadkich czy neologizmów i zapożyczeń, nie oznacza to, że możemy bezkrytycznie podchodzić do jakości tekstów polskojęzycznych produkowanych przez modele.

Najbardziej fundamentalną wadą generatywnej sztucznej inteligencji w kontekście języka polskiego jest nadmiar kalk językowych, spowodowanych negatywnym transferem językowym z angielskiego (tj. bezpośrednim przenoszeniem struktur, zasad lub nawyków z jednego języka do drugiego, które w nowym kontekście stanowią błąd lub owocują nienaturalnymi wypowiedziami). Problem ten może być potęgowany nie tylko przez wykorzystanie w przeważającej ilości danych anglojęzycznych na etapie treningu modelu bazowego, lecz również przez coraz większy udział syntetycznych danych używanych do dostrajania



modelu. Wyobraźmy sobie anglocentryczny model wyjściowy, który szkolimy na rozbudowanym korpusie teksów w języku polskim. Jeśli potem na etapie dostrajania modelu do wykonywania konkretnych zadań, takich jak generowanie e-maili, tekstów piosenek, wniosków i podań czy odpowiadanie na pytania z wiedzy ogólnej, używamy danych destylowanych z innych anglocentrycznych modeli (co jest dość powszechnie przyjętą praktyką) to w podwójny sposób narażamy wrażliwą polszczyznę modelu na kalki i wtrącenia z języka angielskiego. Przejawiają się one w przemycaniu dosłownych fraz charakterystycznych dla angielskiego, a i tak nawet na gruncie tego języka nadużywanych przez modele, np. „Mam nadzieję, że ten e-mail znajdzie Pana w dobrym zdrowiu" – co jest niezgodne z polskimi konwencjami przyjętymi w korespondencji[23].

Dla rzadkich leksemów może się zdarzyć, że na bazie korpusu treningowego modelowi brakuje wystarczającego oparcia, by trafnie przewidzieć następny token (podstawową jednostkę tekstu analizowaną przez model – może to być słowo, jego część lub znak interpunkcyjny). Pojawiają się wtedy mniej lub bardziej absurdalne kontaminacje (połączenia dwóch różnych słów w jedną, nieistniejącą w języku formę), również z użyciem niepolskich znaków alfabetu, jak np. przymiotnik flexibilny. Modelom zdarzają się też, wbrew powszechnym zarzutom o ograniczoną inwencję językową, zabawne, acz nietrafne, tłumaczenia niezaadaptowanych anglicyzmów (np. cyberbullying jako cyberharcowanie). Przykłady z powodzeniem mogłyby wypełniać dziesiątki szpalt czasopism niegdyś znanych jako „humor zeszytów".

Chociaż polszczyzna internetowa, która zapewne stanowi sporą część danych treningowych dla języka polskiego w większości dużych modeli językowych, pozostawia wiele do życzenia, rosnąca obecność tekstów generowanych przez LLM-y potęguje w przestrzeni medialnej również błędną – zwykle nadmiarową – interpunkcję. Szczególnie rzuca się w oczy nagminne stawianie przecinków po frazach inicjalnych w rodzaju „zgodnie z..." czy „ponadto", co odpowiada angielskiej, a nie polskiej normie językowej.

Inną cechą większości najbardziej konkurencyjnych modeli językowych jest tendencja do wielosłowia czy ubierania nawet prostych i krótkich form wypowiedzi w formę mikrorozprawki, podzielonej na wstęp, rozwinięcie i podsumowanie. O ile jednak naszym celem nie jest stworzenie tekstu marketingowego pod wytyczne SEO, niewątpliwie warto popracować nad uszczegółowieniem poleceń (promptów), aby pozyskiwać od modeli bardziej konkretne i pozbawione pustosłowia wypowiedzi.

Prywatność

Stosując narzędzia bazujące na modelach językowych, należy pamiętać, że każde polecenie jest przekazywane na zewnętrzne serwery, często znajdujące się poza granicami Unii Europejskiej w krajach nieobjętych regulacjami RODO. Taki sposób przetwarzania danych wiąże się z poważnymi zagrożeniami dla prywatności użytkownika, takimi jak np. brak gwarancji, że dane będą chronione zgodnie z unijnymi standardami, czy trudnościami z egzekwowaniem praw osoby, której dane dotyczą (np. prawa do usunięcia danych). Wśród innych kluczowych kwestii związanych z wykorzystaniem danych przez SI wymienić można również brak przejrzystości przetwarzania danych (użytkownik powinien wiedzieć, co dzieje się z jego danymi)[24].

Wszystko, co zostanie wpisane przez użytkownika w okienko czatowe, może być następnie analizowane przez dostawcę usługi. Przykładowo firma OpenAI w swoim regulaminie przyznaje, że domyślnie przechowuje historię konwersacji i może wykorzystać ją do dalszego trenowania modeli. Oznacza to, że jeśli dziennikarz/dziennikarka wprowadzi do modelu np. pełny tekst planowanego artykułu śledczego, to te dane mogą trafić na serwery OpenAI i mogą być przeanalizowane przez pracowników firmy lub wykorzystane w kolejnej wersji modelu.

Dodatkowo możliwe jest wydobycie danych treningowych za pomocą specjalnie przygotowanych zapytań[25]. W praktyce oznacza to, że jeśli do trenowania modelu zostały wykorzystane informacje niejawne, inny użytkownik może potencjalnie uzyskać do nich dostęp, zadając odpowiednie pytania.

Zdarzały się już przypadki, w których firmy – jak np. Samsung – zakazywały korzystania z generatywnych modeli językowych po incydentach związanych z wyciekiem danych[26,27]. W jednym z takich przypadków inżynierowie nieświadomie wprowadzili do systemu fragmenty zastrzeżonego kodu źródłowego, który w ten sposób opuścił bezpieczne środowisko firmy[28].

Dlatego zawsze należy unikać umieszczania w zapytaniach kierowanych do modeli językowych danych wrażliwych, poufnych lub takich, które nie powinny wydostać się poza kontrolowane środowisko.

Zasady bezpieczeństwa i dobre praktyki w pracy z LLM

1. Bezpieczeństwo danych osobowych, wrażliwych i niejawnych

Jeśli nie ma możliwości korzystania z bezpiecznego środowiska, czyli modelu lokalnego lub hostowanego na wewnętrznych serwerach firmy, nie należy korzystać z ogólnie dostępnych interfejsów czatowych w celu przetwarzania danych osobowych, wrażliwych czy informacji niejawnych, o których wyciek można się obawiać. Warto przemyśleć anonimizację takich danych przed podaniem ich w poleceniu do modelu. Jeśli rodzaj zadania na to pozwala, można użyć fikcyjnych danych lub zaślepek (np. [nazwisko] lub [kwota]), by wydobyć pożądane odpowiedzi od modelu, a następnie wprowadzić dane faktyczne poza okienkiem czatowym.

2. Weryfikacja treści podawanych przez model

Pod żadnym pozorem nie należy traktować odpowiedzi modelu jako wiedzy bezdyskusyjnej i niepodważalnej, która nie wymaga weryfikacji. Każdy fakt wygenerowany przez model trzeba poddać krytycznej analizie i sprawdzić jego zasadność w zewnętrznych, niezależnych źródłach. Jeśli jakaś informacja, szczególnie związana z niszową czy specjalistyczną tematyką, nie daje się jednoznacznie zweryfikować, nie należy ryzykować jej publikacji bez absolutnej pewności, że ma poparcie w faktach. W szczególności warto zwracać uwagę na podawane przez model cytaty, nazwiska, dane liczbowe czy źródła wiedzy. Ponadto konieczne jest zachowanie ostrożności, jeśli model wyciąga wnioski i problematyzuje relacje między różnymi faktami (stara się je interpretować, komentować, sugerować istnienie zależności etc.). LLM-y, nawet gdy trafnie dokonują ekstrakcji wiedzy z danych treningowych i podają wiarygodne fakty, mogą nie rozumieć występujących między nimi zależności i tym samym halucynować wnioski z rzetelnych danych.

3. Ocena naturalności języka

Wskazane jest zachowanie czujności w zakresie poprawności i naturalności wypowiedzi modelu. Warto upewnić się, że styl i rejestr odpowiadają aktualnym potrzebom oraz że zastosowane słownictwo nie trąci sztucznością, na przykład ze względu na kalki z języka angielskiego. W trosce o wysoką jakość tekstu warto – w razie wątpliwości – sprawdzić, czy użyte związki frazeologiczne są potwierdzone w uznanych słownikach. W przypadku neologizmów i słów rzadkich warto także upewnić się, że model wybrał właściwą formę fleksyjną, a zapis słowa potwierdzony jest w innych źródłach.

4. Stosowanie zasad dobrego promptowania

Przy rozpoczynaniu pracy z modelami językowymi lub w przypadku testowania nowego modelu warto poświęcić trochę czasu na sprawdzenie, jak reaguje

on na poszczególne typy zapytań. Z reguły im więcej konkretów, informacji kontekstowych i jasno sformułowanych wytycznych dotyczących stylu czy formatowania w poleceniu, tym bardziej satysfakcjonująca odpowiedź. Warto zainwestować czas w rozwój swojego własnego stylu w rozmowie z konkretnym modelem językowym, by optymalizować wyniki w kontekście aktualnych potrzeb. Nie należy jednak zapominać, że inne modele mogą reagować odmiennie na najczęściej używane typy zapytań.

5. **Świadomość zagrożeń i edukacja**

   Dzielenie się wiedzą i doświadczeniami w zakresie korzystania z dużych modeli językowych w środowisku zawodowym i prywatnym może przynosić duże korzyści. Im więcej osób będzie zdawało sobie sprawę zarówno z ograniczeń LLM-ów, jak i ich przydatności w konkretnych zadaniach, tym większa będzie efektywność i bezpieczeństwo ich wykorzystania.

6. **Transparentność**

   Wskazane jest oznaczanie treści tworzonych lub współtworzonych przez sztuczną inteligencję, by zachować transparentność wobec odbiorców czy czytelników, w szczególności w zakresie generacji materiałów audiowizualnych.

Mimo długiej listy obowiązkowych zastrzeżeń wobec swobodnego i nieodpowiedzialnego korzystania z modeli językowych ich wartość pozostaje wciąż nieoceniona. Mogą one znacząco przyspieszyć pracę, przetwarzając błyskawicznie dane, oszczędzając żmudnej manualnej dłubaniny czy generując pomysły w momencie, kiedy ludzkiej inteligencji zdarza się cierpieć na twórczy paraliż.

KAROLINA SEWERYN, ANNA KOŁOS, AGNIESZKA KARLIŃSKA

BARTŁOMIEJ DOBOSIEWICZ,
KONRAD MARZĘCKI

Prawne aspekty wykorzystania narzędzi AI do tworzenia treści przez dziennikarzy i innych twórców z branży mediów

03.

Niniejszy artykuł ma na celu zwrócenie uwagi osób zaangażowanych w tworzenie materiałów w szeroko pojętych mediach na wybrane aspekty prawne związane z tym, co określa się zbiorczą nazwą Sztucznej Inteligencji/Artificial Intelligence (AI). AI stanowi aktualnie jedną z najbardziej dynamicznie rozwijających się technologii, czego skutkiem są także nowe wyzwania prawne, w tym coraz liczniejsze głosy dotyczące potrzeby zmian w obecnych przepisach. Z uwagi na szeroki wachlarz problemów i zagadnień związanych z (r)ewolucją AI, zaprezentowanych zostanie kilka wątków wybranych ze względu na ich potencjalny wpływ na branżę mediów. Poniższy tekst nie zawiera porad prawnych, lecz przede wszystkim spostrzeżenia, których celem jest zwrócenie uwagi na istotne problemy i związane z nimi dzisiejsze dylematy. Dla uproszczenia, nawiązania do różnych ról i zawodów, będą dalej skrótowo ujmowane jako odniesienia do „dziennikarza".

W niniejszym rozdziale, poświęconym zagadnieniom prawnym, odstąpiliśmy od stosowania feminatywów. Wynika to ze specyfiki języka aktów prawnych, do których bezpośrednio odwołujemy się w tej części publikacji. W pozostałych rozdziałach stosujemy feminatywy zgodnie z przyjętą praktyką redakcyjną.

Należy zatem zacząć od... przyszłości – czyli unijnego Aktu o Sztucznej Inteligencji („AI Act")[29], który zasadniczo będzie stosowany od 2 sierpnia 2026 r., ale już teraz wyznacza perspektywy mające wpływ na wielu profesjonalistów.

Dlaczego AI Act?

AI Act to rozporządzenie, które jest bezpośrednio stosowane we wszystkich państwach członkowskich Unii Europejskiej. Jego zakres zastosowania jest szeroki i obejmuje dostawców oraz podmioty stosujące AI z UE, ale także spoza UE, jeżeli udostępniają one w UE swoje rozwiązania. Jest to regulacja o charakterze zasadniczym, ale nie „całościowa" i wiele ważnych kwestii, takich jak prawa autorskie czy zasady odpowiedzialności odszkodowawczej pozostanie uregulowanych w innych przepisach. Niezależnie od oceny takiego podejścia, można jednak zakładać, że AI Act będzie stanowić zasadniczy punkt odniesienia dla definiowania AI oraz dla stosowania i interpretowania w UE innych regulacji prawnych w zakresie związanym z AI.

Jak ustalać, czy korzysta się z systemu AI?

AI Act zawiera definicję systemu AI, jednak stopień jej ogólności może w konkretnych przypadkach budzić wątpliwości co do tego, czy korzysta się z systemu AI, czy może z innego rozwiązania. Czym należy się kierować w takiej sytuacji – oprócz oficjalnych wytycznych[30]? AI Act nakazuje dostawcom systemów AI zadbać, aby w przypadkach, gdy systemy takie są przeznaczone do wchodzenia w bezpośrednią interakcję z osobami fizycznymi, osoby takie były informowane o tym, że prowadzą interakcję z systemem AI, chyba że jest to oczywiste dla takiej osoby. Zasadą będzie zatem poleganie na informacji, którą dostawca opatrzy dany system, np. w jego interfejsie czy instrukcji użytkowania. Zalecana jest jednak pewna ostrożność, ponieważ w praktyce będzie możliwy zarówno brak oznaczenia systemu AI przez producenta (przesłanka „oczywistości" interakcji z AI), jak i próba określania mianem systemu AI narzędzi mniej zaawansowanych (niekoniecznie z korzyścią dla ich użytkowników).

Czy AI Act przypisuje dziennikarzowi określoną rolę?

AI Act zawiera definicję tzw. podmiotu stosującego. W zakresie wykonywanej działalności zawodowej jest nim ten, kto wykorzystuje system AI, nad którym sprawuje kontrolę. Wątpliwości może budzić znaczenie frazy „nad którym sprawuje kontrolę", a jej interpretacja może mieć w przyszłości istotne znaczenie. Na przykładzie tzw. „okienek czatowych" może być dyskusyjne to, czy umożliwienie użytkownikowi wysyłania danych do systemu AI (tzw. „input") i otrzymywania w odpowiedzi rezultatów pracy systemu AI (tzw. „output") samo w sobie mogłoby już stanowić „sprawowanie kontroli" nad systemem AI i wydaje się, że należy w szczególności odróżnić kontrolę nad danymi wyjściowymi od kontroli nad samym systemem. Status tzw. użytkownika końcowego w zasadzie

nie został określony w AI Act i w poszczególnych przypadkach może nie być oczywiste, czy działa się jako „podmiot stosujący", personel podmiotu stosującego (np. pracownik redakcji), czy w jeszcze innym charakterze. Jest to natomiast o tyle istotne, że AI Act nakłada obowiązki prawne właśnie na podmioty stosujące.

W praktyce ocena statusu może zatem zależeć od zakresu „kontroli nad systemem", na jaki będzie pozwalać jego licencja, ale też od tego, kto zdecyduje o wykorzystaniu danego systemu AI (np. dziennikarz czy redakcja), kto zaakceptuje jego regulamin lub licencję, a nawet to, czy dziennikarz jest zatrudniony w ramach stosunku pracy, czy świadczy usługi w ramach działalności gospodarczej. Dla bezpieczeństwa warto przyjąć, jako pewne minimum, że w każdym przypadku dziennikarz powinien posiadać odpowiednie kompetencje do używania danego systemu AI, choćby poprzez zapoznanie się z instrukcją obsługi, regulaminem, itp., a w szczególności wskazaniami odnośnie inputów i outputów.

Przejrzystość – czy należy oznaczać artykuły lub ich fragmenty wygenerowane przez AI?

AI Act wprowadza obowiązek, zgodnie z którym podmioty stosujące system AI, który generuje tekst publikowany w celu informowania społeczeństwa o sprawach leżących w interesie publicznym, powinny ujawnić, że został on sztucznie wygenerowany lub zmanipulowany. Obowiązek ten nie będzie jednak mieć zastosowania, w przypadku gdy takie treści zostały następnie poddane weryfikacji przez człowieka lub kontroli redakcyjnej i gdy za publikację treści odpowiedzialność redakcyjną ponosi osoba fizyczna lub prawna. Spełnienie tego warunku pozwoli zatem zwolnić się z obowiązku oznaczenia tekstu jako sztucznie wygenerowanego lub zmanipulowanego.

Zgodnie z innym obowiązkiem, podmioty stosujące system AI, który generuje obrazy, treści audio lub wideo stanowiące deepfake, lub który manipuluje takimi obrazami lub treściami, powinny ujawniać, że dane treści zostały sztucznie wygenerowane lub zmanipulowane. Przewidziano też wyjątek od tej zasady, dotyczący sytuacji, gdy treść stanowi część pracy lub programu o wyraźnie artystycznym, twórczym, satyrycznym, fikcyjnym lub analogicznym charakterze. Obowiązek ogranicza się w takich przypadkach do ujawnienia istnienia takich treści w odpowiedni sposób, który nie utrudnia wyświetlania lub korzystania z utworu.

Które jeszcze przepisy będą istotne dla dziennikarzy korzystających z AI?

W związku z tym, że AI Act nie stanowi regulacji „całościowej", ważne dla dziennikarza zagadnienia będą nadal związane choćby z prawem prasowym, regulacjami dotyczącymi radiofonii i telewizji, reklamy, a także prawem autorskim czy kodeksem cywilnym (np. odpowiedzialność za szkody). W tym kontekście warto przywołać kilka zasad prawa prasowego, których stosowaniu AI wydaje się nadawać dodatkowy kontekst, tj. przepisy dotyczące staranności i rzetelności w działalności dziennikarskiej (art. 12) i tajemnicy dziennikarskiej (art. 15). Dziennikarz jest obowiązany zachować szczególną staranność i rzetelność przy zbieraniu i wykorzystaniu materiałów prasowych, zwłaszcza sprawdzić zgodność z prawdą uzyskanych wiadomości lub podać ich źródło, a także chronić dobra osobiste i interesy działających w dobrej wierze informatorów i innych osób, które okazują mu zaufanie.

Dziennikarz ma też obowiązek zachowania w tajemnicy danych umożliwiających identyfikację autora materiału prasowego, listu do redakcji lub innego materiału o tym charakterze, jak również innych osób udzielających informacji opublikowanych albo przekazanych do opublikowania, jeżeli osoby te zastrzegły nieujawnianie powyższych danych. Dotyczy to także wszelkich informacji, których ujawnienie mogłoby naruszać chronione prawem interesy osób trzecich. W kontekście powyższych przepisów warto przyjrzeć się kilku zagadnieniom wynikającym ze stosowania systemów AI, a związanych z w szczególności z inputem i outputem.

Czy do systemów AI można przesyłać dowolne treści i elementy (input)?

Należy zacząć od tego, że jako input mogą być wprowadzane do systemów AI bardzo różne formy wyrazu, np. tekst (w tym cytat), obraz, dźwięk, etc. Niektóre z nich mogą podlegać ochronie prawnej, np. stanowić utwory, informacje poufne, dane osobowe, tajemnicę zawodową, tajemnicę przedsiębiorstwa. Może zatem powstać ryzyko naruszenia prawa i odpowiedzialność z tym związana (np. karna, cywilna, w tym odszkodowawcza). Takie ryzyko powstanie w szczególności, gdy wśród danych wejściowych znajdą się dane objęte tajemnicą dziennikarską, np. dotyczące informatorów czy materiały pozyskane w ramach śledztwa dziennikarskiego. Przed przesłaniem inputu do systemu AI istotne jest ustalenie, czym może skutkować takie działanie.

Poza tym, że na podstawie inputu system AI wygeneruje output, możliwe jest również wykorzystanie inputu przez dostawcę także do innych celów, np.

do dalszego trenowania systemu AI, czy przekazywania partnerom producenta, co może poskutkować np. wyświetlaniem fragmentów inputu przez system AI innym użytkownikom systemu AI w generowanych dla nich outputach. Aktualna faza rozwoju AI w dużej mierze uzasadnia gromadzenie jak największej ilości danych treningowych, należy się zatem liczyć z powszechnością takiej praktyki. Nie usprawiedliwi to jednak niewystarczającego poziomu ostrożności po stronie wysyłającego input. Przed rozpoczęciem korzystania z systemu AI konieczne jest zatem zapoznanie się z dokumentacją dostawcy – regulaminem, licencją, instrukcją obsługi itp.

Wprowadzenie do systemu AI danych podlegających ochronie. Jakie mogą być konsekwencje?

Jak wcześniej wskazano, niektóre z inputów mogą podlegać ochronie prawnej, np. stanowić informacje poufne, dane osobowe, tajemnicę zawodową, tajemnicę przedsiębiorstwa itd. Ich użycie poprzez wprowadzenie do narzędzia AI rodzi ryzyko naruszenia prawa.

Z punktu widzenia pracy dziennikarza newralgiczna jest odpowiedzialność za: naruszenie tajemnicy dziennikarskiej, naruszenia praw autorskich oraz naruszenia w zakresie przetwarzania danych osobowych.

Po pierwsze, w omawianym kontekście kluczowy jest art. 15 ust. 2 Prawa prasowego, związany z obowiązkiem dochowania tzw. tajemnicy dziennikarskiej. Wprowadzenie do systemu AI inputu zawierającego dane chronione prawem prasowym, może zostać zakwalifikowane jako przestępstwo (występek) określony w art. 266 Kodeksu karnego, zgodnie z którym ten, kto wbrew przepisom ustawy lub przyjętemu na siebie zobowiązaniu, ujawnia lub wykorzystuje informację, z którą zapoznał się w związku z pełnioną funkcją, wykonywaną pracą, działalnością publiczną, społeczną, gospodarczą lub naukową, podlega grzywnie, karze ograniczenia wolności albo pozbawienia wolności do lat 2.

Drugim newralgicznym obszarem w kontekście pracy dziennikarza jest korzystanie z utworów chronionych przepisami prawa autorskiego. Wykorzystanie utworu niezgodnie z warunkami licencji (względnie umowy o przeniesienie autorskich praw majątkowych) może skutkować skierowaniem przeciwko naruszającemu (np. dziennikarzowi) nie tylko roszczeń o charakterze cywilnoprawnym, określonych w prawie autorskim – np. żądania: zaniechania naruszania, usunięcia skutków naruszenia, naprawienia wyrządzonej szkody czy wydania uzyskanych korzyści – ale również zarzutów o charakterze karnym. Najważniejszymi przykładami możliwej odpowiedzialności karnej są: odpowiedzialność za przywłaszczenie sobie autorstwa albo wprowadzanie w błąd co do autorstwa

całości lub części cudzego utworu (art. 155 Prawa autorskiego) oraz za utrwalenie lub zwielokrotnienie cudzego utworu w wersji oryginalnej w celu rozpowszechnienia bez uprawnienia albo wbrew jego warunkom (art. 117 Prawa autorskiego). Dla wszystkich wskazanych wyżej reżimów odpowiedzialności (cywilnej i karnej) zachowaniem, które może prowadzić do jej zdefiniowania, jest właśnie bezprawne wprowadzenie do systemu AI inputu zawierającego utwory chronione prawem autorskim, a następnie wykorzystanie (z naruszeniem warunków korzystania), outputu bazującego na utworze wprowadzonym do takiego systemu.

Trzecim wreszcie kluczowym obszarem, w ramach którego może dojść do naruszenia obowiązujących przepisów prawa, jest przetwarzanie danych osobowych, których przetwarzanie nie jest dopuszczalne albo do przetwarzania których podmiot nie jest uprawniony. W tym zakresie skutkiem naruszenia polegającego na wprowadzeniu do systemu AI inputu zawierającego chronione dane osobowe może być odpowiedzialność karna (art. 107 ustawy o ochronie danych osobowych) lub odpowiedzialność o charakterze cywilnoprawnym, w ramach której każda osoba, która poniosła szkodę majątkową lub niemajątkową w wyniku naruszenia niniejszego rozporządzenia, ma prawo uzyskać od administratora lub podmiotu przetwarzającego odszkodowanie za poniesioną szkodę (art. 82 RODO[31]).

Input – co ważnego dla dziennikarza może wynikać z regulaminu systemu AI?

Dla systemów AI dosyć typowym wymaganiem regulaminowym jest oświadczenie użytkownika co do tego, że przysługują mu odpowiednie prawa do elementów przesyłanych do systemu AI, takich jak utwory, dane osobowe, a także że przesłanie ich nie narusza żadnych prawnie chronionych tajemnic itp. Nie należy takich postanowień traktować jako czegoś „do odkliknięcia", ponieważ mogą nieść daleko idące konsekwencje prawne. Jednocześnie, ich stosowanie trudno traktować jako nadużycie ze strony dostawców. Osoba pracująca w zawodzie dziennikarza powinna zatem ustalić, przed wysłaniem, na jakiej podstawie sama przetwarza dane elementy (np. dla cudzych utworów może to być licencja czy tzw. dozwolony użytek, dla danych osobowych odpowiednia z podstaw wynikających z RODO, itd.) i czy taka podstawa uprawnia ją do ich użycia jako inputu systemu AI.

W odniesieniu do informacji zawartych w inpucie dziennikarz zasadniczo traci faktyczną kontrolę nad nimi już w momencie ich wysłania do systemu AI. Dlatego tak istotny jest zakres przetwarzania, jaki w regulaminie gwarantuje sobie dostawca systemu. Zweryfikowania wymaga rodzaj operacji, którym może

zostać poddany input, zwłaszcza takich, które nie są niezbędne dla wygenerowania outputu (np. użycie do trenowania AI). Istotny jest także zakres odbiorców inputu (np. partnerzy dostawcy), czy lokalizacje, w których możliwe będzie przetwarzanie danych (np. poza EOG). Ma to znaczenie dla wymaganego przez prawo prasowe standardu dbałości o tajemnicę dziennikarską, obowiązków wynikających z RODO czy ochrony dóbr osobistych.

Output – na co warto zwrócić uwagę?

Dziennikarz traci kontrolę nad inputem wysyłając go do systemu AI, natomiast w odniesieniu do outputu sytuacja będzie zasadniczo odwrotna. Materiał przesłany z systemu AI do urządzenia dziennikarza realnie znajdzie się w dyspozycji odbiorcy. Można się z nim zapoznać i podjąć decyzję dotyczącą zakresu i sposobu wykorzystania. Output jest wynikiem działania zautomatyzowanych procesów systemu AI, co jednak nie wyklucza tego, że może obejmować elementy prawnie chronione (np. fragmenty utworów, informacje poufne etc.). Dostawcy systemów AI często zamieszczają w swoich regulaminach klauzule dotyczące „praw do outputu", które zwykle informują o stosunku samego dostawcy do takich rezultatów, jednakże nie muszą przesądzać o charakterze prawnym poszczególnych wyników. Jednocześnie dostawcy mogą wskazywać w regulaminach odbiorcę outputu jako odpowiedzialnego za sposób wykorzystania takiego materiału. Na razie w kwestiach prawnych nie ma prostej odpowiedzi i należy analizować każdy przypadek indywidualnie, w oparciu o zawartość outputu i odpowiednie przepisy, zależnie od tego czy wynik będzie obejmować np. utwór, tajemnicę przedsiębiorstwa, dane osobowe itd.

Analiza prawdziwości informacji jest, w świetle prawa prasowego, podstawowym obowiązkiem dziennikarza. Należy pamiętać, że na razie systemy AI nie dają „gwarancji prawdziwości" i zawsze należy liczyć się z tzw. „halucynowaniem". Takie informacje mogą w dodatku podlegać dalszemu propagowaniu, więc brak należytej weryfikacji outputu może skutkować np. dezinformacją czy naruszeniem dóbr osobistych, a w konsekwencji osłabieniem zaufania do dziennikarza i redakcji.

„Współautorstwo z AI" – czy ma podstawy prawne i co z honorarium autorskim?

Wraz z rozwojem sztucznej inteligencji obserwowana jest praktyka wskazywania tzw. „współautorstwa z AI". Aktualnie prezentowane poglądy dotyczące praw autorskich zasadniczo opierają się na założeniu, że twórcą jest człowiek, a utwory są wynikiem działania ludzi. W tym kontekście uznanie

„współautorstwa" człowieka i systemu AI może rodzić istotne wątpliwości. O ile w ramach takiej „współpracy" nie można wykluczyć możliwości wskazania poszczególnych czynności, które wykonał człowiek, istotną trudność stanowi precyzyjne rozdzielenie „udziału" człowieka i systemu. Może to prowadzić do wątpliwości, czy tak stworzony materiał (tekst, grafika, film itd.) będzie w ogóle korzystać z ochrony prawnoautorskiej, a jeśli tak, to w jakim zakresie.

Twórcy, którzy na podstawie odpowiednich przepisów podatkowych rozliczają preferencyjnie tzw. honorarium autorskie, dokonują tego z uwagi na korzystanie z przysługujących im praw autorskich lub rozporządzanie nimi. Jak wskazał Minister Finansów[32]: „przedmiotem prawa autorskiego jest tylko taka działalność twórcza, która prowadzi do powstania utworu korzystającego z ochrony praw autorskich; sama działalność twórcza nie może być przedmiotem prawa autorskiego". Wątpliwości dotyczące „współautorstwa z AI" mogą więc być dla twórców istotne także w sferze podatków.

„Współpraca z AI" – czy wydawca może sprawdzać dziennikarza?

Rozwój AI umożliwia błyskawiczne gromadzenie danych i istotne skrócenie czasu poświęcanego na opracowanie tekstu. Odpowiednio dobrane i sformułowane prompty pozwalają uzyskać satysfakcjonujące rezultaty. Innymi słowy, AI to narzędzie, które poprzez automatyzację usprawniło proces twórczy. Tak „wspomagana" twórczość może jednak nie zyskać aprobaty wydawców, którzy przed opublikowaniem mają prawo weryfikować czy, dany utwór jest dziełem człowieka. Wdrażają oni w tym zakresie tzw. dobre praktyki, począwszy od odbierania oświadczeń o udziale AI w procesie twórczym, skończywszy na wykorzystaniu odpowiednio wytrenowanych modeli językowych do wykrywania użycia sztucznej inteligencji. Tak, AI bada, czy użyto AI. W procesie wykrywania AI ustala się więc prawdopodobieństwo utworzenia tekstu przez sztuczną inteligencję, np. poprzez badanie charakterystycznych cech modeli językowych, takich jak struktura i przewidywalność tekstu, długość zdań itd. Sztuczna inteligencja, mimo dynamicznego rozwoju, w przeciwieństwie do człowieka nadal „pisze" teksty w sposób bardziej schematyczny i powtarzalny, może mieć trudności z tworzeniem oryginalnych i kreatywnych treści, może „cytować" źródła w sposób mechaniczny i z pominięciem kontekstu, może wreszcie mieć problemy z używaniem idiomów, metafor lub związków frazeologicznych. Stwierdzenie przez wydawcę, że tekst został napisany przy użyciu AI, może być więc podstawą do odrzucenia utworu i, niejako przy okazji, rzutować na wiarygodność osoby, która taki tekst przedstawiła do publikacji jako własny.

Z uwagi na format poradnika, niniejszy artykuł zasygnalizował jedynie wybrane zagadnienia w formule hasłowej. Istotna część aktualnych wątpliwości prawnych związana jest z coraz szerzej sygnalizowaną potrzebą dostosowania przepisów do wyzwań, które niesie AI. Poruszone kwestie mogą ulegać dynamicznym zmianom oraz podlegać nowym regulacjom, wobec czego należy traktować je jako spostrzeżenia według stanu „na dziś" i stosować do nich podejście krytyczne, podobnie jak do outputu systemów AI (choć sam artykuł został stworzony bez udziału AI).

BARTŁOMIEJ DOBOSIEWICZ, KONRAD MARZĘCKI

SYLWIA ADAMCZYK

ABC zagrożeń w przestrzeni informacyjnej. Kluczowe pojęcia i strategie przeciwdziałania dezinformacji

04.

Dynamiczny rozwój technologii wpływa na zacieranie granic między prawdą a fikcją – fałszywe treści stają się coraz trudniejsze do odróżnienia od wiarygodnych informacji. Użytkownicy internetu codziennie poruszają się w nieskończonym strumieniu wiadomości. W natłoku napływających treści, to media stają się jednym z ważniejszych ogniw w ochronie społeczeństwa przed dezinformacją.

Weryfikacja informacji w erze algorytmów i automatyzacji

Według Reuters Institute Digital News Report 2025[33], najczęściej wybieranym źródłem wiadomości w Polsce jest internet (portale internetowe i aplikacje, media społecznościowe, agregatory newsów, oraz inne platformy cyfrowe) – korzysta z niego 77% respondentów. Wyprzedza on telewizję (54%) i prasę papierową (10%).

Ten sam raport pokazuje, że same media społecznościowe są wykorzystywane do pozyskiwania wiadomości przez niemal połowę Polaków (48%). Tymczasem to właśnie na platformach społecznościowych każdy może publikować dowolne informacje bez ich wcześniejszej weryfikacji, a fałszywe wiadomości rozprzestrzeniają się szybciej niż prawdziwe – co pokazały analizy badaczy z Massachusetts Institute of Technology w 2018 roku. Wykazano w nich, że fałszywe informacje potrzebowały sześć razy mniej czasu na dotarcie do użytkowników i miały średnio 70% więcej szans na udostępnienie na ówczesnym Twitterze (dziś X), niż informacje prawdziwe[34].

Platformy społecznościowe są projektowane tak, żeby jak najdłużej zatrzymać uwagę użytkowników – spędzony na korzystaniu z nich czas bezpośrednio przekłada się na zyski ich właścicieli. Dlatego algorytmy analizują, które treści wywołują najwięcej polubień, komentarzy czy udostępnień. Treści sensacyjne i wzbudzające silne emocje naturalnie generują więcej takich reakcji, dlatego są premiowane przez algorytmy. Sprzyja to rozprzestrzenianiu się nieprawdziwych lub zmanipulowanych informacji, które w dużej mierze bazują na emocjach. Dodatkowo personalizacja strumienia aktualności sprawia, że użytkownicy zamykają się w tzw. bańkach informacyjnych, gdzie widzą głównie treści zgodne z ich własnymi poglądami. Utrudnia to poznanie innych opinii czy perspektyw i zniekształca obraz rzeczywistości.

W tych okolicznościach to właśnie profesjonalne media, w których istnieją mechanizmy kontroli i weryfikacji, pełnią kluczową rolę w dostarczaniu odbiorcom rzetelnych informacji. Weryfikacja treści w redakcjach odgrywa tutaj fundamentalną rolę, szczególnie w obliczu współczesnej dynamiki środowiska informacyjnego. Tempo pracy dziennikarzy i priorytet szybkości przekazu sprawiają, że skrupulatna weryfikacja faktów staje się jeszcze bardziej istotna – choć również bardziej wymagająca, zwłaszcza w kontekście rozwoju technologii umożliwiających aktorom dezinformacji zautomatyzowaną produkcję i masową dystrybucję fałszywych treści.

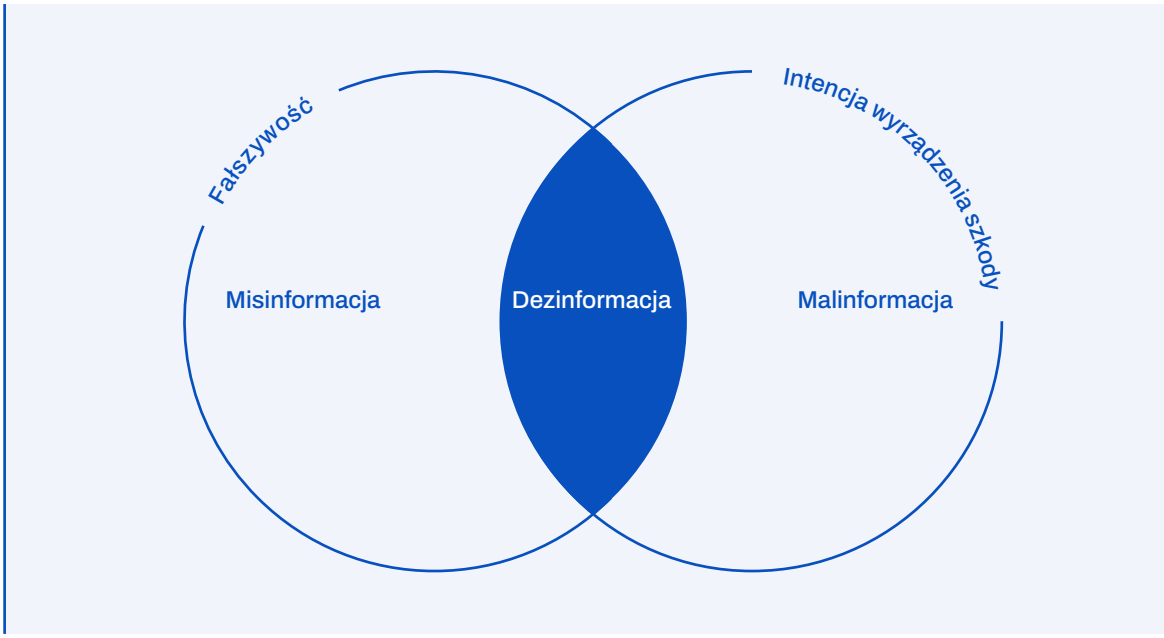
Podstawowe definicje i wybrane techniki manipulacji

Znajomość podstawowych terminów związanych z zagrożeniami w przestrzeni informacyjnej pozwala lepiej rozumieć, jak działają mechanizmy manipulacji. Pomaga świadomie reagować na próby wprowadzenia w błąd, a także skuteczniej identyfikować zagrożenia i rozpoznawać intencje nadawców. Na początek warto zatem rozróżnić pojęcia takie jak dezinformacja, misinformacja i malinformacja.

- **DEZINFORMACJA**
  Fałszywa bądź wprowadzająca w błąd treść, która jest tworzona lub rozpowszechniana celowo, z zamiarem oszukania odbiorców bądź pozyskania określonych korzyści: np. ekonomicznych czy politycznych.

- **MISINFORMACJA**
  Fałszywa bądź wprowadzająca w błąd treść, która jest tworzona bądź rozpowszechniana bez złych zamiarów oraz bez świadomości jej nieprawdziwości – w wyniku pomyłki lub błędnego przekonania na dany temat. Kluczową różnicą między dezinformacją i misinformacją jest intencja.

- **MALINFORMACJA**
  Rodzaj zakłócenia w sferze informacyjnej, w którym wykorzystywane są prawdziwe treści, ale w złej intencji – z zamiarem wyrządzenia szkody. Może to być ujawnienie poufnych informacji w domenie publicznej, publikowanie bądź rozpowszechnianie treści prywatnych czy intymnych, bez zgody osoby lub instytucji, których dotyczą.



RYS. 1. Schemat zaburzeń informacyjnych[35]

W przekazach dezinformacyjnych stosowane są różnorodne techniki manipulacji mające na celu zniekształcenie rzeczywistości i modelowanie opinii odbiorców według zamierzeń nadawcy. Wśród takich metod można wymienić między innymi:

- **Podszywanie się** – używanie logotypu lub wizerunku osoby lub organizacji w celu wykorzystania ich wiarygodności do szerzenia szkodliwych informacji.
- **Dowód anegdotyczny** – argument opierający się na osobistych doświadczeniach lub pojedynczych przykładach, niepoparty badaniami lub danymi statystycznymi. Prowadzi do błędnych wniosków, zakładając przeniesienie indywidualnego doświadczenia na ogólny trend.
- **Manipulowanie danymi** – wyciąganie fałszywych wniosków z danych, ich niepoprawna interpretacja lub wycinanie z kontekstu w celu wprowadzenia w błąd.
- **Cherry-picking** – wykorzystywanie wybiórczych dowodów lub danych na potwierdzenie określonej tezy, przy jednoczesnym ignorowaniu pozostałych materiałów, które przeczyłyby temu twierdzeniu.

- **Fałszywy kontekst** – przedstawianie informacji w ujęciu, które wprowadza odbiorcę w błąd.
- **Fałszywe powiązanie** – tworzenie pozornych zależności między informacjami w rzeczywistości niezwiązanymi ze sobą, w celu wprowadzenia odbiorcy w błąd poprzez wywołanie wrażenia, że dwa czynniki są od siebie zależne.
- **Teoria spiskowa** – alternatywne wyjaśnienie zdarzenia, zakładające istotny udział grupy konspiratorów, próbujących zataić prawdę przed opinią publiczną.

Świadomość opisanych technik pomaga skuteczniej rozpoznawać potencjalnie fałszywe lub zmanipulowane treści. Stanowi też podstawową ochronę przed nieświadomym powielaniem takich praktyk we własnej pracy.

Przeciwdziałanie dezinformacji

Choć skala wyzwań jest znacząca, istnieją sprawdzone strategie umożliwiające przeciwdziałanie omawianym zagrożeniom. Należą do nich debunking, prebunking oraz fact-checking. Każde z tych podejść odgrywa istotną rolę w ograniczaniu wpływu dezinformacji.

- ### DEBUNKING I PREBUNKING

Debunking (z ang. „obalanie", „demaskowanie") to działanie polegające na ujawniania fałszu i manipulacji. Jego celem jest ograniczenie wpływu nieprawdziwych i potencjalnie szkodliwych informacji. Debunking dotyczy konkretnych treści, które zdążyły już zyskać popularność.

Innymi słowy, debunking polega na ujawnianiu fałszywych informacji i obalaniu ich przy użyciu wiarygodnych źródeł, po tym jak odbiorcy zetkną się z dezinformacją.

Prebunking natomiast polega na proaktywnym ostrzeganiu osób przed dezinformacją jeszcze przed jej napotkaniem, obalaniu często powtarzających się błędnych argumentów oraz wyjaśnianiu metod manipulacji stosowanych w rozpowszechnianiu fałszywych informacji. Idea prebunkingu polega więc na przygotowaniu ludzi na kontakt z fałszywymi informacjami.

Kampanie prebunkingowe, mające na celu budowanie odporności społeczeństwa na fałszywe przekazy, są najczęściej prowadzone przez instytucje i organizacje zajmujące się przeciwdziałaniem dezinformacji.

Badania z 2024 roku wskazują[36], że prebunking i debunking stanowią skuteczne i komplementarne strategie przeciwdziałania dezinformacji. Badacze podkreślają, że efektywność obu metod zależy od kontekstu: prebunking skuteczniej buduje odporność na techniki manipulacji, podczas gdy debunking sprawdza się w korygowaniu fałszywych tez. Obie strategie mogą być stosowane równolegle – prebunking jako działanie prewencyjne i debunking jako reakcja na konkretne fałszywe narracje.

- **FACT-CHECKING**

Fact-checking (z ang. „weryfikacja faktów") jest procesem weryfikacji informacji w celu sprawdzenia ich prawdziwości. Fact-checking można przeprowadzić zarówno przed publikacją informacji, w celu zweryfikowania zawartych w niej danych – co powinno stanowić dobrą praktykę w procesie tworzenia każdego materiału – jak i po jej opublikowaniu, co z kolei jest domeną analityków pracujących na rzecz instytucji, organizacji czy serwisów fact-checkingowych.

W badaniu PressInstitute z 2022 roku[37], zatytułowanym „Problem dezinformacji w ocenie polskich dziennikarzy", niemal połowa respondentów (46%) przyznała, że ich pracodawca „nie zabezpiecza dziennikarzy, nie narzuca rozwiązań, sami dbają o weryfikację informacji". 29% dziennikarzy zadeklarowało, że pracodawca „sformułował wewnętrzne procedury weryfikacji informacji zabezpieczające dziennikarzy". 15% wskazało, że redakcja posiada „własny dział fact-checkingu". 7% nie wie, jakie działania w tym zakresie podejmuje pracodawca, z kolei 3% respondentów stwierdziło, że redakcja współpracuje w tym obszarze z podmiotem zewnętrznym. W opinii większości osób biorących udział w badaniu, procedura fact-checkingu powinna odbywać się poprzez wewnętrzny dział w każdej redakcji (61%). Respondenci wskazują, że oprócz weryfikacji informacji takie działy powinny zajmować się m.in. także budowaniem społecznej świadomości dotyczącej fałszywych treści, ostrzeganiem przed potencjalnie szkodliwymi narracjami oraz edukacją.

Brak systemowego wsparcia ze strony redakcji w obszarze fact-checkingu może zwiększać ryzyko misinformacji, zwłaszcza w warunkach presji czasowej i intensywnego tempa pracy. Przywołane dane wskazują zatem na potrzebę budowania redakcyjnych standardów weryfikacji treści oraz inwestowania w kompetencje, które pozwolą dziennikarzom skuteczniej przeciwdziałać manipulacjom w infosferze.

Dezinformacja jako wyzwanie współczesności

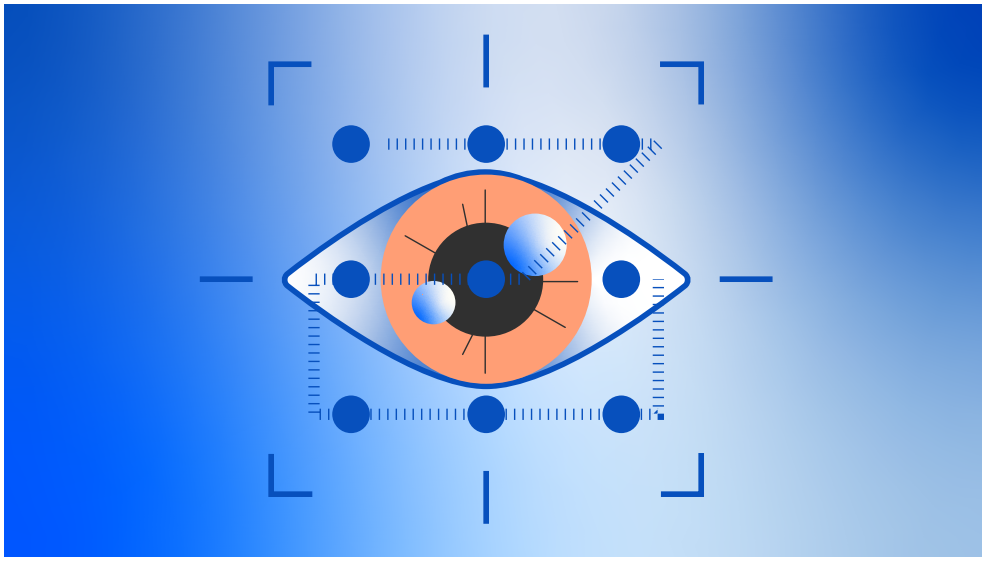
W 2025 roku dezinformacja i misinformacja drugi raz z rzędu zostały uznane za najpoważniejsze globalne zagrożenie w perspektywie krótkoterminowej, zajmując pierwsze miejsce w rankingu ryzyk Światowego Forum Ekonomicznego „The Global Risks Report”[38]. Dezinformacja rozprzestrzenia się obecnie szybciej niż kiedykolwiek wcześniej. Na tę sytuację ma wpływ m.in. dynamiczny rozwój generatywnej sztucznej inteligencji i automatyzacja produkcji fałszywych treści. W tej sytuacji dziennikarze, którzy dostarczają odbiorcom rzetelne i sprawdzone wiadomości, odgrywają wyjątkowo istotną rolę w ochronie przestrzeni informacyjnej.

SYLWIA ADAMCZYK

EWELINA BARTUZI-TROKIELEWICZ,
ALICJA MARTINEK

Deepfake: zagrożenia dla środowiska informacyjnego

05.



Dezinformacja to mechanizm, który z precyzją celuje w naszą świadomość, jest oparty na celowych działaniach oraz zaburzaniu przekazu i służy osiąganiu korzyści. Stanowi zagrożenie, z którym media muszą aktywnie walczyć. To właśnie pracownicy mediów odgrywają kluczową rolę w identyfikowaniu i demaskowaniu manipulacji. Dynamiczny rozwój sztucznej inteligencji przyniósł liczne korzyści, wspierając dziennikarzy w analizie danych, automatyzacji pracy redakcyjnej i weryfikacji źródeł. Jednak te same narzędzia, które ułatwiają życie, niosą także liczne wyzwania.

Technologia, która zmienia świat

Jednym z najbardziej niepokojących zastosowań sztucznej inteligencji jest technologia generatywna, pozwalająca tworzyć realistyczne, a jednocześnie fałszywe treści. Technologie generatywne, w szczególności bazujące na manipulacjach materiałami audiowizualnymi – tzw. deepfake – redefiniują zaufanie do obrazu i dźwięku. Jeszcze kilka lat temu nagranie wideo było niemal równoznaczne z dowodem. Dzisiaj tak naprawdę każdy może wygenerować przekonujący materiał, w którym ktoś wygłasza oświadczenie, którego nigdy nie wypowiedział, albo rekomenduje fałszywą inwestycję.

Deepfake umożliwia syntetyczne odwzorowanie twarzy, głosu czy mimiki. Materiały te są masowo rozpowszechniane w sieci, często jako reklamy, treści viralowe lub posty o charakterze dezinformacyjnym. Mogą one nie tylko zmylić odbiorców, ale także wpływać na postawy społeczne, decyzje konsumenckie, a nawet wyniki wyborów. Co gorsza, generowanie takich materiałów stało się

bardzo proste. Obecnie nie wymaga dużej wiedzy, zasobów obliczeniowych ani specjalistów – wystarczy dostęp do internetu i podstawowe umiejętności obsługi aplikacji, by w ciągu kilku minut stworzyć fałszywy przekaz.

Jak deepfake zagraża środowisku informacyjnemu?

Technologia generatywna znajduje także wiele pozytywnych zastosowań – od rozrywki i edukacji, przez wsparcie procesów śledczych, poszukiwanie osób zaginionych, aż po ochronę tożsamości osób znajdujących się w sytuacjach zagrożenia. Natomiast w rękach nieodpowiednich osób staje się bronią o bardzo wysokim potencjale niszczenia reputacji, siania chaosu i podważania wiarygodności informacji. Analiza przypadków wykorzystywania syntetycznych materiałów wskazuje na różnorodne cele, jakie realizują ich twórcy. Są to m.in.:

- **Fałszywe reklamy promujące przede wszystkim nieistniejące produkty finansowe** – platformy inwestycyjne, aplikacje do szybkiego zarabiania pieniędzy, rzekome rabaty, dopłaty rządowe, programy państwowe, a także niesprawdzone lub fikcyjne środki medyczne i terapie. Czasami odnoszą się także do rzekomych zmian prawnych lub edukacyjnych bądź wprowadzania nowych zakazów. Często wykorzystują wizerunki polityków, dziennikarzy, ekspertów w danej dziedzinie, celebrytów, a nawet duchownych, by nadać przekazowi autorytet i zwiększyć jego wiarygodność.
- **Manipulacja treścią przemówień politycznych** – zmiana wypowiedzi znanych osób w celu szerzenia dezinformacji lub wzbudzenia kontrowersji.
- **Syntetyczne materiały audio i video do wyłudzania danych** – voice cloning i podszywanie się pod znajomych w komunikatorach („na wnuczka”, „na policjanta”, a teraz też „na przyjaciela”).
- **Materiały o charakterze dyskredytacyjnym i kompromitującym** – przedstawiające ludzi w stanie upojenia, w sytuacjach o zabarwieniu erotycznym/pornograficznym (tzw. deepnude) czy jako ofiary przemocy.
- **Zastraszanie** – inscenizacje pobić, symulacje śmierci, groźby wobec dziennikarzy mające zmusić ich do działań (np. usunięcia materiału).
- **Fałszywe oferty pomocy** – wykorzystywane w oszustwach finansowych, nierzadko jako „druga fala” ataku na osoby wcześniej oszukane, powołujące się na międzynarodowe pseudoinstytucje, zniechęcające do zgłaszania oszustw na policję i coraz częściej wykorzystujące wizerunki funkcjonariuszy.
- **Satyra i memy polityczne** – nie zawsze szkodliwe, ale mogą prowadzić do dezinformacji, gdy nie są jasno oznaczone jako humorystyczne.

- **Fikcyjne profile trolli i farmy lajków** – wykorzystujące generowane obrazy i teksty do tworzenia zasięgu, testowania zachowania użytkowników i późniejszej manipulacji.
- **Fałszywe nagrania wydarzeń** – symulacje katastrof, defilad wojskowych czy zjawisk paranormalnych, które mogą wpływać na emocje społeczne, a nawet rynki finansowe.
- **Wprowadzanie w błąd osób publicznych przez tzw. pranki (żarty polegające na celowym wciąganiu nieświadomej ofiary w niezręczną sytuację)** – wymierzone w polityków lub celebrytów, mogące prowadzić do kompromitacji lub utraty reputacji.
- **Zastosowania w procesach sądowych** – materiały deepfake są wykorzystywane jako rzekome dowody w sprawach sądowych lub jako strategia obrony polegająca na podważeniu prawdziwości materiału („to deepfake").

W środowisku informacyjnym coraz częściej mówi się o zjawisku tzw. „efektu fałszywego odwrotu" (ang. liar's dividend), polegającego na tym, że równie groźne jak akceptowanie fałszywych materiałów jest odrzucanie prawdziwych jako rzekomych treści deepfake.

Wizerunki dziennikarzy w oszustwach: jak i dlaczego?

Zagrożenia dla mediów w kontekście deepfejków obejmują przede wszystkim utratę zaufania odbiorców. Coraz więcej osób zaczyna podważać autentyczność nawet prawdziwych materiałów. Z drugiej strony, utratą wiarygodności

może skutkować nawet przypadkowe udostępnienie fałszywych materiałów pod presją czasu czy pogonią za reakcjami.

Warte podkreślenia są także celowe ataki na dziennikarzy i redakcje, gdzie wizerunki znanych medialnie osób są wykorzystywane do oszustw, np. w reklamach inwestycyjnych, farmaceutycznych czy politycznych (rys. 2).

RYS. 2. Przykładowy kadr z fałszywej reklamy wykorzystujący wizerunek prezenterki Moniki Majewskiej z dnia 24.03 2023 roku.

Deepfejki wykorzystujące twarze i głosy dziennikarzy są szczególnie skuteczne, bo bazują na zaufaniu. Natomiast wykorzystanie szablonu serwisu informacyjnego zwiększa wiarygodność przekazu (sugerując, że coś było rzekomo pokazane w telewizji). Ofiary oszustw otrzymują reklamy z pozornie autentycznymi wypowiedziami znanych prezenterów czy reporterów. Typowe scenariusze obejmują polecanie nowej platformy inwestycyjnej przez dziennikarza/dziennikarkę, zapowiedź innej osoby, która prezentuje szczegóły oferty, informowanie o rewolucyjnym leku ukrywanym przez koncerny.

Dlaczego to działa? Ponieważ dla większości odbiorców osoby medialne są autorytetami. Dodatkowo wykorzystanie wizualnego formatu „newsowego" czy z programów śniadaniowych dodaje powagi i jednocześnie zwiększa wiarygodność przekazu. Oszuści wiedzą, że deepfejki wzbudzają emocje i działają szybciej niż sam tekst. Na ich korzyść działa jeszcze ludzka psychologia – ofiary

oszustw często są przekonane, że dowiadują się o „tajemnicy" przeznaczonej dla „wybranych", co tylko sprzyja podejmowaniu impulsywnych decyzji.

Konsekwencje płynące z rozprzestrzeniających się materiałów deepfake to między innymi naruszenie dobrego imienia. A co za tym idzie – potencjalne sprawy sądowe. Wraz z naruszeniem reputacji dziennikarzy, społeczeństwo może tracić zaufanie do instytucji medialnych. Tempo rozsiewania deepfejków w sieci często może wywierać presję na działy PR i zespoły wideo lub redakcyjne. Potrzeba bardzo szybkiego ustosunkowania się do danych treści podnosi ryzyko popełnienia błędu. Problemy te nie są wizją z przyszłości, ponieważ wielu dziennikarzy w Europie Środkowej zgłaszało już przypadki kontaktu z deepfejkami poprzez biblioteki reklam platform społecznościowych (publicznie dostępne bazy danych, które pozwalają użytkownikom sprawdzić aktualnie emitowane na danej platformie reklamy). Fałszywe treści są tam dostosowywane do lokalnych warunków, często z użyciem imienia i nazwiska osoby publicznej, zwiększając szansę na rozpoznanie danej osoby i zadziałanie jej autorytetu.

Jak rozpoznać deepfake bez narzędzi?

Choć jakość deepfejków stale rośnie, zwłaszcza dzięki wykorzystaniu coraz lepszych narzędzi komercyjnych, większość materiałów nadal zawiera typowe błędy, które mogą zostać zauważone przez uważnych odbiorców. Rozpoznawanie deepfejków wymaga czujności, doświadczenia i znajomości typowych błędów, jakie popełnia sztuczna inteligencja. Część z nich jest widoczna dopiero w zwolnionym tempie lub analizie poklatkowej.

Wśród charakterystycznych artefaktów można wyróżnić:

- **Artefakty na krawędziach twarzy i przy przesłonięciach** – na styku twarzy z otoczeniem często pojawiają się charakterystyczne błędy: podwójne kontury, dodatkowe elementy, obszarowe łączenie obrazów, nierówne brzegi czy „rozlanie" skóry poza naturalną linię. Szczególnie widoczne są one, gdy twarz jest częściowo zasłonięta dłonią, włosami, mikrofonem lub innym obiektem (rys. 3).
- **Nienaturalne rozmycia i kontrasty między twarzą a tłem** – w przypadku niedokładnie przygotowanych deepfejków, często pojawia się tzw. efekt maski, gdzie twarz wygląda jak sztucznie nałożona na ciało. Wynika to z braku post-processingu, czyli z niedopasowania koloru skóry, kontrastu i tekstury między nakładaną twarzą a szyją lub resztą ciała z materiału źródłowego (prawdziwego). Twarz może być zbyt jasna, zbyt gładka lub wyglądać jak doklejony element.

RYS. 3. Przykłady artefaktów w okolicy twarzy. Od lewej: dublowanie konturów twarzy, błędy przy przesłonięciu twarzy dłonią oraz mikrofonem.

- **Niespójna mimika i błędy synchronizacji ust** – ruchy ust mogą być opóźnione względem dźwięku lub nieadekwatne do wypowiadanych słów. Mimika często nie oddaje emocji, wygląda nienaturalnie lub jest ograniczona do minimalnych ruchów.
- **Artefakty w okolicy ust** – najczęstsze błędy możliwe do wizualnego wykrycia obejmują obszar ust. Może być nienaturalnie rozmyty lub nadmiernie wyostrzony w porównaniu do reszty twarzy, co wynika z trudności modeli generatywnych z realistycznym odwzorowaniem dynamicznej struktury ust podczas mowy. Zęby bywają przedstawione w sposób zdeformowany – mogą być nieproporcjonalne, źle osadzone względem linii ust lub wygenerowane z niewystarczającą szczegółowością. Błędy te są związane z brakiem precyzji w renderowaniu anatomicznych detali.
- **Brak synchronizacji mowy ciała z wypowiedzią** – w sytuacji niedopasowania materiału źródłowego do spreparowanego przekazu ruchy głowy, rąk i ciała mogą być niespójne z tonem i treścią wypowiedzi, np. ktoś mówi emocjonalnie, ale pozostaje całkowicie nieruchomy.
- **Wygładzona tekstura twarzy** – niektóre metody, które generują twarze, a nie modyfikują istniejące materiały, charakteryzują się nienaturalnie gładką, pozbawioną detali skórą. W takich materiałach brakuje elementów charakterystycznych dla realistycznego obrazu, m.in. porów, zmarszczek, przebarwień, asymetrii.
- **Dodatkowe lub zniekształcone elementy ciała** – algorytmy mogą mieć trudność z odwzorowaniem szczegółów ciała i wynikowe materiały mogą zawierać błędy w anatomii, np. nadmiar palców, nienaturalne zgięcia dłoni, asymetryczne kończyny czy niewłaściwe połączenia stawów.

- **Niespójności szczegółów w tle** – mimo iż głównym celem manipulacji są zazwyczaj wizerunki osób, czasami pojawiają się zniekształcenia w tle: powielone przedmioty, krzywe linie, rozmyte napisy, artefakty, błędne odbicia w lustrze lub nieciągłości w oświetleniu i perspektywie
- **Zmieniony akcent, skoki głośności, metaliczny pogłos** – w syntezowanej mowie mogą występować nienaturalne przejścia między akcentami: osoba mówiąca po polsku nagle przechodzi w akcent innego języka, np. rosyjski lub angielski, co jest efektem działania modeli wielojęzykowych. Dodatkowo słyszalne mogą być wyraźne zmiany w natężeniu dźwięku (nagłe ściszenia lub podbicia) oraz charakterystyczny metaliczny pogłos, typowy dla niskiej jakości syntezy głosu.
- **Maszynowe tempo** – syntetyczna mowa często cechuje się nienaturalnym, w miarę jednostajnym tempem, przypomina czytanie, z krótkimi odstępami między wyrazami, bez przyspieszeń czy o obniżonej intonacji. Takie „robotyczne" brzmienie jest jednym z najbardziej rozpoznawalnych sygnałów fałszywego głosu.
- **Słyszalne oddechy** – głos może zawierać także nienaturalne przerwy oraz cykliczne i wyraźne oddechy.
- **Błędy wymowy** – w syntetycznych wypowiedziach często pojawiają się pomyłki fonetyczne, czyli różnego rodzaju zniekształcenia polskich głosek, błędne wymawianie słów, w szczególności nazw własnych lub liczebników, a także niepoprawne ich odmiany. W niektórych przypadkach brzmienie jest tak nienaturalne, że trudno nawet zapisać je fonetycznie, wypowiedź „rozjeżdża się" względem znanych wzorców językowych i brzmi obco lub niezrozumiale.
- **Błędy językowe, tłumaczeniowe i gramatyczne** – treści generowane z użyciem modeli językowych mogą zawierać błędy składniowe, niepoprawne tłumaczenia oraz kalki językowe. Często pojawiają się też niezgodności gramatyczne, niepoprawne odmiany wyrazów lub błędna składnia zdań.
- **Logiczne niespójności** – związane z niedopasowaniem płci/tożsamości, zawierające niezgodność z przepisami prawa i/lub oczekiwaniami społecznymi, niespójności liczbowe czy semantyczno-pragmatyczne.

> „Jeśli w **wieku dwóch tysięcy dwudziestu czterech lat** będziesz nadal leczyć stawy pigułkami...”

> „Zapomniałem o ciężkości i bólu trzustki, **a koledzy mówią mi, że wrócił mi wzrok**.”

> „Wstyd się przyznać, ale sam miałem prostatę i teraz dzięki naszemu produktowi czuję sie o **dwadzieścia lat starszy.**”

Dla ukrycia tych błędów twórcy materiałów stosują techniki takie jak:

- **Stosowanie szumów teksturalnych** (rys. 4) – w celu ukrycia nienaturalnych cech wygenerowanego obrazu, takich jak zbyt gładka skóra, zniekształcenia wokół oczu czy ust, twórcy deepfejków wprowadzają do obrazu tzw. szumy teksturalne. Mogą to być lekkie ziarnistości, symulacje zakłóceń transmisji, ziarno filmowe, maski teksturalne, znaki wodne lub filtry graficzne. Takie zabiegi „zanieczyszczają” obraz w sposób, który utrudnia zarówno ludzkie rozpoznanie fałszerstwa, jak i wykrycie przez modele uczone na czystych danych syntetycznych.

RYS. 4. **Tekstury wykorzystywane przez oszustów by ukryć deepfejki. Od lewej: fraktale, efekt social media, „brudna szyba”, orientalna tekstura, rozmycie, kropki.**

- **Dodawanie szumów w ścieżce dźwiękowej** (melodie, hałasy uliczne, śmiech, głosy) – aby utrudnić wykrycie manipulacji, do sklonowanego głosu coraz częściej celowo dodawane są tła dźwiękowe, które maskują nienaturalności

w mowie syntetycznej i mogą zakłócać działanie modeli wykrywających manipulacje. Melodie, szumy uliczne, śmiech czy przypadkowe głosy skutecznie zagłuszają niedoskonałości generatorów, utrudniając zarówno ocenę autentyczności przez ludzi, jak i automatyczne analizy.

- **Zmiana sposobu prezentacji materiału przez nagrywanie ekranu z telefonu lub laptopa** – manipulacja prezentowana jest jako nagranie z ekranu innego urządzenia. Ta technika obniża jakość w sposób naturalny, np. dodając rozmycia, odbicia światła, perspektywę czy zniekształcenia, pozwalając ukryć typowe artefakty deepfejków. Dla systemów wykrywających materiały syntetyczne oznacza to zmianę tzw. domeny danych i znaczne utrudnienie analizy.
- **Umieszczanie syntetycznego materiału w długich neutralnych segmentach** – fragmenty deepfejka są wkomponowywane w długie, prawdziwe materiały, np. w neutralne filmy przyrodnicze lub prezentowanie statycznego obrazu przez kilkadziesiąt minut. Dzięki temu syntetyczny materiał jest ukryty wśród prawdziwych danych, zmniejszając prawdopodobieństwo zauważenia go przez administratorów lub automatyczne systemy kontroli danych.
- **Stosowanie kolaży, efektów postprodukcyjnych i montażu** – techniki obejmujące m.in. cięcia, przejścia, kolaże czy filtry graficzne, mają za zadanie celowo zakrywać błędy typowe dla deepfejków lub odwracać od nich uwagę.

Typologie materiałów syntetycznych

Na podstawie analizy incydentów można wyróżnić kilka głównych typów materiałów:

- **Jednolite deepfejki (ok. 31%)** – materiały, w których przez całe trwanie filmu, kadr pozostaje niezmieniony. Możliwe są cięcia, jednak cały klip przedstawia konsekwentnie jedną osobę, najczęściej w formie jej monologu.
- **Materiały złożone (ok. 60%)** – które obejmują kompilację wielu źródeł, gdzie zmanipulowany materiał jest przeplatany prawdziwymi filmami, materiałami stockowymi (pobranymi z banków zdjęć i filmów), wypowiedziami wielu osób. Do tej kategorii należą materiały charakteryzujące się złożonym montażem – na przykład wywiady z osobami o wygenerowanym wizerunku lub nagrania przedstawiające jedną osobę zarejestrowaną z różnych ujęć kamery.
- **Kolaże (ok. 19%)** – kompleksowe montaże, które mają różne formy, przyjmują m.in. postać rozmowy wielu ludzi, imitując wideorozmowy albo rozmowy z ekspertami. Zdarza się także umieszczanie mówiących osób na tle jakiejś wizualizacji. To najbardziej zaawansowana montażowo grupa deepfejków. Materiały takie często zawierają „stockowe” wstawki i tak np. w materiale opowiadającym o inwestowaniu w gazociąg mogą pojawić się

ujęcia prezentujące rury w polu. Taki sposób prezentowania treści, często idzie w parze z wizualnym udawaniem serwisów informacyjnych, których charakterystyczna formuła ma za zadanie dodatkowo uwiarygodnić prezentowane treści. Kolaże towarzyszą zarówno deepfejkom jednolitym (jako „okno w oknie" podczas stałego kadru na jedną osobę), jak i złożonym.

Pod względem zawartości audio wyróżniamy:

- DF-A (93%) – pełna manipulacja dźwiękiem (voice cloning).
- Partial deepfake (7%) – mieszanie elementów prawdziwych i syntetycznych.
- Pełne deepfake audio-wideo (69%) – zarówno obraz, jak i dźwięk są sztucznie generowane.
- Częściowe manipulacje obrazu (4%) lub użycie wyłącznie materiałów stockowych.

Złożoność i socjotechnika – jak działają kampanie z wykorzystaniem deepfejków?

Wprowadzające w błąd reklamy oparte na deepfejkach coraz częściej przybierają postać złożonych kampanii medialnych. W ich realizacji wykorzystywane są sceny wideo, animacje, obrazy, a także materiały generowane syntetycznie. Często spotykaną praktyką jest łączenie fragmentów z prawdziwych serwisów informacyjnych z treściami zmanipulowanymi lub wygenerowanymi – takimi jak fałszywe wywiady z lekarzami, artystami, politykami, ekspertami lub rzekome wypowiedzi znanych osób. Dla zwiększenia realizmu twórcy wykorzystują również treści pornograficzne, nagrania postronnych osób oraz zasoby stockowe.

Charakterystycznym elementem takich materiałów jest obecność transkrypcji wypowiadanego tekstu – nadaje im wiarygodności w sposób typowy dla treści publikowanych w mediach społecznościowych. W celu zwiększenia perswazyjności stosowane są zaawansowane techniki modyfikacji mimiki i głosu przy użyciu uczenia maszynowego (lip-sync – z ang. lip synchronization, czyli „synchronizacja ust").

Całość przekazu jest starannie skonstruowana w oparciu o mechanizmy manipulacyjne, w tym:

- budowanie presji czasu („oferta ograniczona czasowo"),
- wzbudzanie strachu („proszę o natychmiastowe zamknięcie wszystkich aptek"),
- personalizacja komunikatów („jestem taki jak Ty"),

- fałszywe recenzje i komentarze rzekomych użytkowników („wywiady”, chmurki wiadomości),
- powoływanie się na autorytety – lekarzy, duchownych, profesorów,
- narracje spiskowe i ujawnienie „zakazanej” wiedzy.

Z analiz wynika, że grupami, których wizerunek najczęściej bywa wykorzystywany w takich kampaniach, są:

- przedstawiciele władz państwowych i ich rodziny – politycy, ministrowie, urzędnicy, sędziowie,
- prezesi spółek, właściciele firm, biznesmeni,
- artyści – piosenkarze, aktorzy, reżyserzy, pisarze, sportowcy, celebryci, influencerzy,
- redaktorzy, dziennikarze, twórcy medialni,
- lekarze, naukowcy, komentatorzy i eksperci.

Grupy docelowe – potencjalne ofiary manipulacji

Przekazy wykorzystujące technologię generatywną są często kierowane do określonych grup społecznych, szczególnie podatnych na manipulację.

Najczęściej są to:

- osoby chore, szczególnie zmagające się z chorobami przewlekłymi i cywilizacyjnymi,
- osoby starsze, nieposiadające wiedzy technicznej,
- osoby skłonne do inwestowania oszczędności,
- osoby już poszkodowane w przeszłości, szukające pomocy,
- młodzież i dzieci,
- osoby podatne na manipulację, sugestię, działające impulsywnie.

Złożoność formy przekazu, jego audiowizualna atrakcyjność oraz wykorzystanie znanych twarzy i głosów sprawiają, że materiały te są bardzo skuteczne. Dodatkowo, kampanie są często targetowane – dopasowywane do preferencji i historii wyszukiwań odbiorcy/odbiorczyni. Różne grupy społeczne mogą otrzymywać całkowicie odmienne przekazy deepfejkowe, ale każdy z nich

będzie maksymalnie dostosowany do konkretnego odbiorcy/odbiorczyni poprzez język i wykorzystanie dobranych autorytetów oraz dostosowanie formy przekazu.

Wyzwania dla mediów i dziennikarstwa

Media pełnią kluczową rolę w demokratycznym społeczeństwie, dostarczając informacji, które kształtują nasz obraz świata. Zaufanie społeczne do mediów opiera się na przekonaniu, że prezentowane wiadomości są rzetelne i starannie weryfikowane. Jednak w obliczu rosnącej dezinformacji to zaufanie wystawione jest na coraz większą próbę. W tej sytuacji szczególnego znaczenia nabierają następujące wyzwania:

- **Wzmacnianie etyki dziennikarskiej** – odpowiedzialność za przekaz, weryfikacja źródeł, transparentność procesów redakcyjnych i unikanie stronniczości.
- **Unikanie selektywności poznawczej** – ważne jest prezentowanie informacji w pełnym kontekście, bez pomijania istotnych faktów. Redakcje powinny świadomie przeciwdziałać zjawisku tzw. cherry-pickingu, czyli wybierania tylko tych danych, które potwierdzają z góry założoną tezę.
- **Systematyczna weryfikacja informacji** – deepfejki i inne formy dezinformacji bazują na emocjach. Ważne jest systematyczne sprawdzanie każdej informacji niezależnie od formy i unikanie nieświadomego szerzenia nieprawdziwych treści.
- **Rozwijanie kompetencji technicznych** – redakcje powinny inwestować w narzędzia i szkolenia pomagające wykrywać treści syntetyczne. Wskazana jest również współpraca z ekspertami technologicznymi.
- **Edukowanie społeczeństwa** – odbiorcy muszą być świadomi istnienia deepfejków, umieć je identyfikować i rozumieć ich wpływ. Media pełnią istotną rolę w podnoszeniu świadomości społecznej. Promowanie krytycznego myślenia, edukowanie w zakresie rozpoznawania manipulacji oraz uświadamianie istnienia fałszywych treści wzmacnia odporność społeczną na dezinformację oraz na wynikające z niej szkody.

Media, jako filar demokratycznego społeczeństwa, muszą mieć możliwość identyfikowania fałszywych treści i zapewniania rzetelności informacji. Tylko wtedy będą w stanie skutecznie pełnić swoją misję – nie tylko informacyjną, ale również edukacyjną i kontrolną.

Jak sobie radzić z deepfejkami?

W erze syntetycznej rzeczywistości zaufanie stanie się kluczowym zasobem. Redakcje, które postawią na edukację, weryfikację i transparentność, mają szansę nie tylko przetrwać, ale i zyskać przewagę. Deepfejki nie znikną – ale możemy nauczyć się je rozpoznawać i demaskować. Warto też pamiętać, że fałszywy materiał może szkodzić nie tylko reputacji redakcji, ale i realnym osobom. Dlatego by walczyć z deepfejkami kluczowa jest edukacja – nie tylko profesjonalistów, ale także dzieci i seniorów, którzy ze względu na mniejsze doświadczenie technologiczne są szczególnie podatni na manipulację. Na poziomie firm i instytucji ważne jest, by wdrażać procedury zgłaszania incydentów i zapewnić dostęp do szkoleń z analizy fałszywych materiałów.

Musimy być świadomi, że walka z deepfejkami wymaga skoordynowanej współpracy pomiędzy działami funkcjonującymi w redakcji, platformami mediów społecznościowych, działami prawnymi i IT, a także w niektórych przypadkach instytucjami takimi jak CERT Polska czy Policja.

Istnieją też organizacje fact-checkinkgowe i debunkingowe, które posiadają usługę ekspertyzy analizy materiałów syntetycznych. Warto się z nimi kontaktować, ponieważ deepfejki często trafiają do nich różnymi kanałami, zwiększając szansę na to, że dany materiał został już przeanalizowany.

Gdzie zgłaszać incydenty?

W sytuacji dyskredytacji lub bezprawnego użycia wizerunku osoby prywatnej najlepiej dokonać zgłoszenia na policję, ponieważ inne jednostki, jak np. CERT, nie mają narzędzi do ścigania przestępców. Dobrą praktyką jest również zgłaszanie deepfejków do administratorów serwisów, gdzie takie materiały zostały opublikowane. Dodatkowo – na adres deepfake@nask.pl można wysyłać materiały do weryfikacji. Deepfejki, których zadaniem często jest wyłudzanie danych, można zgłaszać na incydent.cert.pl. Natomiast materiały pornograficzne, w szczególności dotyczące nieletnich, najlepiej kierować bezpośrednio do Dyżurnetu (dyzurnet.pl/zglos-nielegalne-tresci). W rzeczywistości, w której prawda staje się towarem deficytowym, naszym wspólnym zadaniem jest dbanie o jej wartość.

Co w przypadku stania się ofiarą?

W przypadku stania się ofiarą podobnego incydentu nie należy zachowywać bierności. Warto zgłosić sprawę na policję. Tylko poprzez zwracanie uwagi organów ścigania na skalę problemu można oczekiwać stworzenia skuteczniejszych metod walki z przestępczością internetową.

Ważne jest, by mówić o tym, co się wydarzyło. Warto wykorzystać dostępne kanały komunikacji, aby otwarcie poinformować, że dany materiał jest nieprawdziwy, dane słowa nigdy nie zostały wypowiedziane, a postać występująca w filmie została sztucznie wygenerowana.

Warto skorzystać z prawa do zapomnienia. Obywatelom Unii Europejskiej przysługuje możliwość żądania usunięcia powiązania pomiędzy fałszywym materiałem a nazwiskiem. Można to osiągnąć, zwracając się do wyszukiwarek internetowych z prośbą o usunięcie połączenia pomiędzy tożsamością a niepożądanymi treściami. Jest to szczególnie przydatne w sytuacji, gdy źródło fałszywych informacji znajduje się na nieetycznych stronach.

EWELINA BARTUZI-TROKIELEWICZ, ALICJA MARTINEK

MICHAŁ MAREK, OLGA WAŁKUSKA,
PATRYCJA KRZYŚPIAK, KORNEL KOWIESKI

Ingerencje podmiotów zewnętrznych w polską infosferę

06.

Rosyjska i białoruska dezinformacja jest w Polsce zjawiskiem znanym, badanym i szeroko opisywanym. Zainteresowanie tym tematem znacząco wzrosło od czasu rosyjskiej inwazji na Ukrainę. Choć świadomość społeczną można określić jako wysoką, to jednak należy pamiętać, że Rosjanie nieustannie doskonalą swój aparat dezinformacyjny.

Obok znanych oficjalnych kanałów propagandowych, takich jak Sputnik czy RT (Russia Today) coraz większe znaczenie zyskują zdecentralizowane i często anonimowe źródła wpływu w mediach społecznościowych, zwłaszcza na platformie Telegram i X (dawniej Twitter). W polskojęzycznej infosferze funkcjonuje szereg kont i kanałów, które stanowią element tzw. rosyjsko-białoruskiego łańcucha informacyjnego.

Przekaz dezinformacyjny prowadzony przez te źródła charakteryzuje się cechami opisanymi poniżej:

- **Unikanie bezpośrednich odniesień do Federacji Rosyjskiej** – zamiast wychwalać Rosję, pisze się o „antysystemowym realizmie”, „prawdziwym patriotyzmie”, „suwerenności” lub „tradycyjnych wartościach”.
- **Użycie prawdziwych informacji, które są wyjęte z kontekstu lub zmanipulowane** – cytaty lub prawdziwe wydarzenia przedstawiane są w kontekście, który wypacza, a czasem zupełnie zmienia ich wydźwięk i znaczenie.
- **Działanie w ramach tzw. klastrów narracyjnych** – można wyróżnić główne tematy, wokół których budowane są konkretne „opowieści”, np. „zdrada Zachodu”, prześladowania prorosyjskich dziennikarzy, rusofobia jako narzędzie cenzury, NATO jako okupant.

Klaster narracyjny to zestaw przekazów, komunikatów lub narracji, które różnią się szczegółami, ale koncentrują się wokół jednej tezy. Jego cechami są: spójność tematyczna, wykorzystanie różnych kanałów oraz sposobów docierania do obiorców i rozpowszechniania przekazu (artykuły, wpisy w mediach społecznościowych oraz siatki botów), odwoływanie się do silnych emocji.

- **Używanie wizerunku Polaków lub osób polskiego pochodzenia jako nośników przekazu** – informacje pochodzące od Polaków wzbudzają dużo mniej podejrzeń. Podobnie takie, które pisane są poprawną polszczyzną.
- **Wykorzystywanie aktualnych napięć społeczno-politycznych** – im bardziej polaryzujący temat (migracja, relacje polsko-ukraińskie, spory polityczne), tym lepiej. Przekazy bazujące na emocjach są częściej udostępniane, zatem mają szanse dotrzeć do większej liczby odbiorców.

Ogólny zarys działania mechanizmu propagandy

Treści zgodne z interesami geopolitycznymi Kremla i Mińska publikowane są w polskojęzycznych serwisach, które pozycjonują się jako niezależne i opiniotwórcze. Następnie narracje te pojawiają się na platformie Telegram w formie cytatów, zrzutów ekranu czy streszczeń, a także stają się inspiracją dla wpisów użytkowników X, YouTube'a oraz alternatywnych blogów. Warto zaznaczyć, że na platformie Telegram oraz w serwisie X działają siatki kont zajmujące się powielaniem i udostępnianiem takich przekazów w celu zwiększenia ich zasięgów.

Wybrane taktyki rosyjskiej i białoruskiej dezinformacji w polskojęzycznym internecie

- **Segmentacja przekazu** – treści są dopasowywane do grupy docelowej (np. skrajna prawica, skrajna lewica, środowiska antyszczepionkowe, aktywiści antyunijni).
- **Techniki audiowizualne** – w treściach dezinformacyjnych wykorzystuje się zmanipulowane filmy, deepfejki oraz atrakcyjne graficznie posty.
- **Wzmacnianie emocji** – narracje wykorzystują obawy związane z wojną, kryzysem ekonomicznym, migracją i zmianami kulturowymi.
- **Atakowanie instytucji** – narracje mają za zadanie podważać zaufanie do Sejmu, rządu, służb specjalnych, UE i NATO.

Najważniejsze informacje

W polskojęzycznej przestrzeni internetowej, szczególnie na platformie Telegram i w serwisie X, prowadzone są prorosyjskie i probiałoruskie operacje dezinformacyjne.

Operacje te są coraz trudniejsze do wykrycia, ponieważ nie są prowadzone wprost, zamiast tego odwołują się do idei takich jak: wolne media, niezależne opinie lub obrona wartości narodowych.

Treści są często dostarczane przez polskojęzycznych użytkowników lub osoby ukrywające tożsamość.

Chińska propaganda w polskiej infosferze

W ostatnich latach w polskich mediach dużo uwagi poświęcano dezinformacji i propagandzie rosyjskiej i białoruskiej. Jednocześnie temat podobnych przekazów docierających do polskiej infosfery ze źródeł związanych z Chińską Republiką Ludową nie jest w naszym kraju przedmiotem medialnej debaty.

Należy jednak podkreślić, że Chińska Republika Ludowa prowadzi intensywną współpracę z Federacją Rosyjską na płaszczyźnie politycznej, gospodarczej, militarnej oraz informacyjnej. Rosyjska agencja TASS I jej chiński odpowiednik Xinhua współpracują co najmniej od 2017 r. Współpracę kilkukrotnie odnawiano przy okazji spotkań przedstawicieli Rosji i Chin. W dniach 16 – 17 maja 2024 r. W. Putin odwiedził X. Jinpinga w Pekinie, a w oświadczeniu wydanym po spotkaniu informowano m.in. o kontynuowaniu współpracy w dziedzinie wymiany informacji. Na podstawie tych ustaleń rosyjskie kanały medialne legalnie działają w chińskich mediach społecznościowych, podobnie jak konta niektórych rosyjskich polityków. Rosyjskie kanały są obecne w chińskich mediach, a płynące z nich przekazy mogą, za pośrednictwem aplikacji takich jak TikTok czy RedNote, przenikać do europejskiej infosfery.

Temat chińskiej dezinformacji i propagandy jest mało znany Polakom przekonanym, że kraj nieposiadający z Polską historycznych konfliktów (co nie jest prawdą, gdyż Chińczycy nie zapominają o uznaniu przez IIRP istnienia

marionetkowego Cesartwa Mandżukuo[39]) i odległy geograficznie nie stanowi żadnego zagrożenia. Jednocześnie Chiny, które jeszcze niedawno kojarzone były z komunizmem, biedą i łamaniem praw człowieka, obecnie odbierane są jako mocarstwo rozwinięte technologicznie, gospodarczo i kulturowo. Temat autorytarnego systemu politycznego schodzi na drugi plan.

Chińska propaganda promowana jest poprzez następujące rodzaje źródeł:

1. Kanały chińskie

To m.in. strony i kanały telewizji CGTN i Chińskiego Radia Międzynarodowego (oba podmioty działają w ramach China Media Group) na Facebooku, X czy YouTubie, a także profile działające na platformie Facebook oznaczone jako „media kontrolowane przez państwo" (np. „Studio Słonecznik"), czy też strony internetowe udające polskie profile informacyjne. Skupiają się na rozpowszechnianiu tzw. „przekazu tygodnia" bezpośrednio z Departamentu Propagandy, ale mogą też okazjonalnie zamieszczać informacje zbieżne z rosyjską propagandą i dezinformacją.

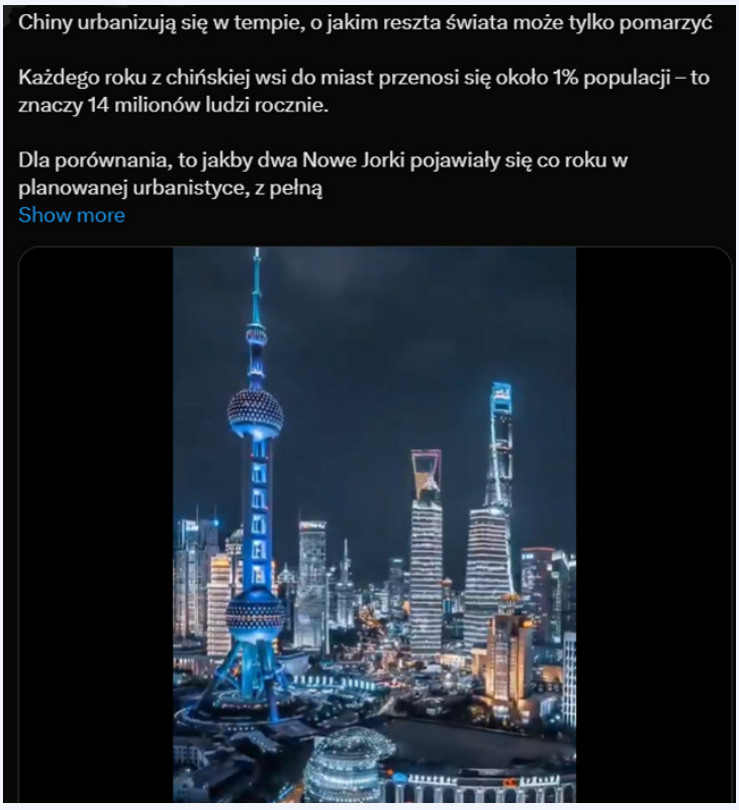
Nie stanowią one w takiej formie poważnego zagrożenia, gdyż ich przekaz jest nadmiernie propagandowy lub pisany przy użyciu translatora, w związku z czym nie trafia do przeciętnego polskiego odbiorcy. Emitowana treść może jednak z czasem ulec profesjonalizacji.

RYS. 5. Profil powiązany z China Media Group promuje nazwę Xizang zamiast Tybet, co stanowi przykład działań na rzecz sinizacji.

2. Polskie kanały powiązane z Chinami

W polskiej infosferze działa szereg kanałów prowadzonych przez osoby posiadające prywatne, biznesowe bądź akademickie powiązania z Chińską Republiką Ludową. Niektóre z nich bezpośrednio współpracują z podmiotami chińskimi (na przykład z China Media Group), a co do niektórych istnieje ryzyko współpracy z chińskimi służbami bezpieczeństwa. Promują one głównie narracje o wyższości Chin nad państwami zachodu w dziedzinie technologii, gospodarki i kultury. Rozpowszechniają też przekaz o potędze Armii Ludowo-Wyzwoleńczej, często wraz z narracją o „jednych Chinach" i nieuchronnym „powrocie" Tajwanu do ChRL. Nierzadko zamieszczają materiały wideo z chińskich platform (np. Douyin, chińskojęzyczna wersja TikToka) lub powołują się na anglojęzyczne chińskie gazety (People's Daily, Global Times). Lobbują też za zbliżeniem Polski (oraz całej UE) z Chinami i prowadzą działania na rzecz pogorszenia publicznej percepcji Unii Europejskiej i NATO.

Stanowią największe zagrożenie, gdyż przedstawiają przekaz propagandowy w sposób atrakcyjny dla odbiorcy, często pod pozorem treści „lifestylowych" lub opinii eksperckich.



RYS. 6. Przykład z wysokozasięgowego kanału rozpowszechniającego prochińską propagandę, materiał z Douyin w tle.

3. Nieświadome powielanie propagandy

W polskich mediach można też zaobserwować nieświadome powielanie propagandy ChRL, często wynikające z chęci stworzenia popularnych materiałów lub odniesienia korzyści materialnych. Portale o tematyce motoryzacyjnej regularnie zamieszczają duże ilości materiałów mających promować chińską motoryzację. W podobny sposób promuje się turystykę do Chin, gdzie pod pozorem zachwalania wybranych atrakcji przekazuje się elementy chińskiej propagandy. Należy zwrócić uwagę na problem zamieszczania w polskich mediach bezpośrednich przedruków z chińskiej prasy (głównie anglojęzycznej). Najczęściej cytowanym źródłem jest South China Morning Post, który jako anglojęzyczny dziennik z Hong Kongu, jest często uważany za obiektywne źródło informacji na temat Chin. W rzeczywistości gazeta od 2015 roku należy do chińskiego koncernu Alibaba Group Holding Limited, który posiada szereg powiązań z KPCh[40] i stanowi narzędzie *soft power* rządu w Pekinie właśnie dzięki pozornej niezależności.

RYS. 7. Przykład promocji chińskiej motoryzacji na portalu o nowych technologiach. Część z artykułów to bezpośrednie przedruki z SCMP.

Wraz z wybuchem wojny na Ukrainie oraz zmianą rozkładu sił w regionie, a także w związku ze zmianami w polityce USA wobec Europy, rząd w Pekinie intensyfikuje działania mające na celu pozyskanie jak największych wpływów na Starym Kontynencie. W strategicznym interesie Polski leży zachowanie ostrożności przy powielaniu informacji z kraju posiadającego tak silne związki z Federacją Rosyjską. W szczególności należy być wyczulonym na przekazy pochodzące bezpośrednio od osób posiadających silne związki z Chinami lub mieszkających tam na stałe. Treści publikowane przez obcokrajowców, nawet w zachodnich mediach, również znajdują się pod kontrolą aparatu państwowego, tak więc zawarty w nich przekaz zawsze będzie promował narracje zbieżne z interesami Chin.

MICHAŁ MAREK, OLGA WAŁKUSKA, PATRYCJA KRZYŚPIAK, KORNEL KOWIESKI

Wykaz źródeł

1 PAP, *Cyberatak na PAP. W serwisie Polskiej Agencji Prasowej nieprawdziwa depesza o mobilizacji,* https://www.pap.pl/aktualnosci/cyberatak-na-pap-w-serwisie-polskiej-agencji-prasowej-nieprawdziwa-depesza-o [dostęp: 22.05.2025 r.]

2 PAP, *Wicepremier Gawkowski: zabezpieczamy PAP przed cyberatakami w przyszłości*, https://www.pap.pl/aktualnosci/wicepremier-gawkowski-zabezpieczamy-pap-przed-cyberatakami-w-przyszlosci [dostęp: 24.06.2025 r.]

3 Bezpieczny Miesiąc, *Kopia zapasowa – Twoja cyfrowa polisa bezpieczeństwa*, https://bezpiecznymiesiac.pl/bm/aktualnosci/1461,Kopia-zapasowa-Twoja-cyfrowa-polisa-bezpieczenstwa.html [dostęp: 23.05.2025r.]

4 L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu (2025). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, https://doi.org/10.1145/3703155 [dostęp: 24.06.2025 r.]

5 K. Emslie (2024). *LLM Hallucinations: A Bug or A Feature?*, https://cacm.acm.org/news/llm-hallucinations-a-bug-or-a-feature/ [dostęp: 24.06.2025 r.]

6 H. Zhang, S. Diao, Y. Lin, Y. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, T. Zhang (2024). *R-Tuning: Instructing Large Language Models to Say 'I Don't Know'*. W: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, https://aclanthology.org/2024.naacl-long.394/ [dostęp: 24.06.2025 r.]

7 The Guardian (2025). *Norwegian files complaint after ChatGPT falsely said he had murdered his children*, https://www.theguardian.com/technology/2025/mar/21/norwegian-files-complaint-after-chatgpt-falsely-said-he-had-murdered-his-children [dostęp: 24.06.2025 r.]

8 Reuters (2025). *Trouble with AI hallucinations spreads to big law firms*, https://www.reuters.com/legal/government/trouble-with-ai-hallucinations-spreads-big-law-firms-2025-05-23/ [dostęp: 24.06.2025 r.]

9 D. Barman, Z. Guo, O. Conlan (2024). *The dark side of language models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination*, https://www.researchgate.net/publication/378878897_The_Dark_Side_of_Language_Models_Exploring_the_Potential_of_LLMs_in_Multimedia_Disinformation_Generation_and_Dissemination [dostęp: 24.06.2025 r.]

10 S. B. Shah, S. Thapa, A. Acharya, K. Rauniyar, S. Poudel, S. Jain, A. Masood, U. Naseem (2024). *Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords*, https://www.researchgate.net/publication/380997232_Navigating_the_Web_of_Disinformation_and_Misinformation_Large_Language_Models_as_Double-Edged_Swords [dostęp: 24.06.2025 r.]

11 C. Chen, K. Shu (2024). *Combating misinformation in the age of LLMs: Opportunities and challenges*, https://doi.org/10.1002/aaai.12188 [dostęp: 24.06.2025 r.]

12 D. Macko, A. A. Ramakrishnan, J. S. Lucas, R. Moro, I. Srba, A. Uchendu, D. Lee (2025). *Beyond speculation: Measuring the growing presence of LLM-generated texts in multilingual disinformation*, https://arxiv.org/abs/2503.23242 [dostęp: 24.06.2025 r.]

13 M. Wack, C. Ehrett, D. Linvill, P. Warren (2025). *Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign*, https://doi.org/10.1093/pnasnexus/pgaf083 [dostęp: 24.06.2025 r.]

14 C. Chen, K. Shu (2023). *Can LLM-generated misinformation be detected?*, https://arxiv.org/abs/2309.13788 [dostęp: 24.06.2025 r.]

15 M. Sadeghi, I. Blachez (2025). *A well-funded Moscow-based global ‚news' network has infected Western artificial intelligence tools worldwide with Russian propaganda*, https://www.newsguardrealitycheck.com/p/a-well-funded-moscow-based-global [dostęp: 24.06.2025 r.]

16 The Washington Post (2025). *LLM Poisoning: How Russia is Grooming Chatbots to Spread Disinformation*, https://www.washingtonpost.com/technology/2025/04/17/llm-poisoning-grooming-chatbots-russia/ [dostęp: 24.06.2025 r.]

17 Benchmark.pl (2025). *Rosja przejmuje internet? Narzędzia AI powtarzają propagandę*, https://www.benchmark.pl/aktualnosci/rosja-przejmuje-internet--narzedzia-ai-powtarzaja-propagande.html [dostęp: 24.06.2025 r.]

18 Tagesspiegel Background (2025). *KI-Hintertür in die Demokratie*, https://background.tagesspiegel.de/it-und-cybersicherheit/briefing/ki-hintertuer-in-die-demokratie [dostęp: 25.07.2025 r.]

19 R. Navigli, S. Conia, B. Ross (2023). *Biases in Large Language Models: Origins, Inventory, and Discussion*, https://doi.org/10.1145/3597307 [dostęp: 24.06.2025 r.]

20 I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed (2023). *Bias and Fairness in Large Language Models: A Survey*, https://doi.org/10.48550/ARXIV.2309.00770 [dostęp: 24.06.2025 r.]

21 PLLuM (2024). *Tworzenie modeli za pomocą modeli, czyli wartość oryginalnych zbiorów danych uczących*, https://pllum.org.pl/blog/posts/tworzenie-modeli-za-pomoca-modeli [dostęp: 24.06.2025 r.]

22 The Guardian (2024). *TechScape: How cheap, outsourced labour in Africa is shaping AI English*, https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt [dostęp: 24.06.2025 r.]

23 PLLuM (2024). *Tworzenie modeli za pomocą modeli*, https://pllum.org.pl/blog/posts/tworzenie-modeli-za-pomoca-modeli [dostęp: 25.06.2025 r.]

24 I. Barberá (2025). *AI Privacy Risks & Mitigations for Large Language Models*, https://www.edpb.europa.eu/our-work-tools/our-documents/support-pool-experts-projects/ai-privacy-risks-mitigations-large_en [dostęp: 25.06.2025 r.]

25 N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, C. Raffel (2021). *Extracting training data from large language models*, https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting [dostęp: 25.06.2025 r.]

26 Business Insider (2023). *Samsung bans employees from using AI tools like ChatGPT and Google Bard after an accidental data leak, report says*, https://www.businessinsider.com/samsung-chatgpt-bard-data-leak-bans-employee-use-report-2023-5 [dostęp: 25.06.2025 r.]

27 The Wall Street Journal (2023). *Apple Restricts Use of ChatGPT, Joining Other Companies Wary of Leaks*, https://www.wsj.com/tech/apple-restricts-use-of-chatgpt-joining-other-companies-wary-of-leaks-d44d7d34 [dostęp: 25.06.2025 r.]

28 Business Insider (2023). *Samsung bans employees from using AI tools like ChatGPT and Google Bard after an accidental data leak, report says*, https://www.businessinsider.com/samsung-chatgpt-bard-data-leak-bans-employee-use-report-2023-5 [dostęp: 25.06.2025 r.]

29 Komisja Europejska (2025). *Regulatory framework on Artificial Intelligence*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [dostęp: 11.06.2025 r.]

30 Komisja Europejska (2024). *Commission publishes Guidelines on AI system definition to facilitate first AI Act's rules application*, https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application [dostęp: 11.06.2025 r.]

31 Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych). Dz.U. UE L 119 z 4.5.2016, s. 1–88, https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A32016R0679 [dostęp: 27.06.2025 r.]

32 Podatki.gov.pl (2020). *Interpretacja ogólna – 50% kup*, https://www.podatki.gov.pl/pit/wyjasnienia-pit/interpretacja-ogolna-50-kup/ [dostęp: 11.06.2025 r.]

33 Reuters Institute (2025). *Digital News Report 2025*, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025 [dostęp: 27.06.2025 r.]

34 S. Vosoughi, D. Roy, S. Aral (2018). *The spread of true and false news online*, https://www.science.org/doi/10.1126/science.aap9559 [dostęp: 20.05.2025 r.]

35 C. Wardle, H. Derakhshan (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*, https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html [dostęp: 20.05.2025 r.]

36 H. Bruns, F.J. Dessart, M. Krawczyk, S. Lewandowsky, M. Pantazi, G. Pennycook, P. Schmid, L. Smillie (2024). *Investigating the role of source and source trust in prebunks and debunks*, https://www.nature.com/articles/s41598-024-71599-6 [dostęp: 20.05.2025 r.]

37 PressInstitute (2022). *Problem dezinformacji w ocenie polskich dziennikarzy*, https://www.press.pl/tresc/72639,nowe-badanie-pressinstitute_-redakcje-musza-uczyc-walki-z-fake-newsami [dostęp: 26.05.2025 r.]

38 World Economic Forum (2025). *The Global Risks Report 2025*, https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf [dostęp: 26.05.2025 r.]

39 E. Pałasz-Rutkowska (2006). *Polska – Japonia – Mandżukuo. Sprawa uznania Mandżukuo przez Polskę*, https://www.academia.edu/105425583/Polska_Japonia_Mand%C5%BCukuo_Sprawa_uznania_Mand%C5%BCukuo_przez_Polsk%C4%99 [dostęp 21.05.2025]

40 Propastop.org (2019). *The Chinese propaganda machine: The awakening giant turns to the West*, https://www.propastop.org/en/2019/05/15/the-chinese-propaganda-machine-the-awakening-giant-turns-to-the-west/ [dostęp: 21.05.2025 r.]

NASK
Ministerstwo Cyfryzacji
PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI