ROCZNIKI NAUK SPOŁECZNYCH
Tom 16(52), numer 1 – 2024
DOI: https://doi.org/10.18290/rns2024.0009

JAN KREFT
MONIKA BOGUSZEWICZ-KREFT
BARBARA CYREK

HALUCYNACJE CHATBOTÓW A PRAWDA: GŁÓWNE NURTY DEBATY I ICH INTERPRETACJE

CHATBOT HALLUCINATIONS VS. TRUTH: MAINSTREAM DEBATES AND THEIR INTERPRETATIONS

A b s t r a c t. Generative artificial intelligence (AI) systems are able to create media content by applying machine learning to large amounts of training data. This new data may include text (e.g. Bard by Google, LLaMa by Meta, or ChatGPT by OpenAI) and visuals (e.g. Stable Diffusion or DALL-E by OpenAI) and audio (e.g. VALL-E by Microsoft). The level of advancement of this content may make it indistinguishable from human creativity. However, chatbots are characterized by the so-called hallucinations, which are largely a new type of disinformation. The aim of the research undertaken is to identify the main trends in the debate on the effects of the use of artificial intelligence, with particular emphasis on disinformation involving chatbots in the media environment. The study adopted the research method of a systematic literature review, which, among other things, limits selection bias. The interpretation of the main trends in the debate leads to the conclusion that the disinformation of chatbots in the form of their hallucinations is significant in terms of scale, is optimized and personalized, and has a significant potential to erode social trust.

Keywords: chatbots; disinformation; artificial intelligence.

Prof. dr hab. JAN KREFT – Politechnika Gdańska, Wydział Zarządzania i Ekonomii, Centrum Badań nad Zarządzaniem Algorytmicznym, Zakład Zarządzania Algorytmicznego; adres do korespondencji: ul. Gabriela Narutowicza 11/12, 80-233 Gdańsk; e-mail: jankreft@pg.edu.pl; ORCID: https://orcid.org/0000-0003-4129-8424.

Dr hab. MONIKA BOGUSZEWICZ-KREFT, prof. WSB Merito – Uniwersytet WSB Merito w Gdańsku, Wydział Biznesu, Katedra Marketingu; adres do korespondencji: al. Grunwaldzka 238A, 80-266 Gdańsk; e-mail: monika.boguszewicz@gdansk.merito.pl; ORCID: https://orcid.org/0000-0002-9294-7175.

Dr BARBARA CYREK – Uniwersytet Dolnośląski DSW Wrocław; adres do korespondencji: Strzegomska 55, 53-611 Wrocław; e-mail: cyrek.barbara@gmail.com; ORCID: https://orcid.org/0000-0002-3270-6548.

WPROWADZENIE

O ile przełom lat 2022/2023 był początkiem masowej popularności chatbotów, to w kolejnych latach należy się spodziewać upowszechnienia sztucznej inteligencji w codziennym życiu, czyli jej obecności w różnego rodzaju urządzeniach: od smartfonów po sprzęty powszechnego użytku, od okularów sterowanych głosem po urządzenia gospodarstwa domowego. Tymczasem nasilająca się konkurencja na rynku chatbotów oznacza rozwój złożonych systemów sztucznej inteligencji, które będą mogły reagować na kombinację danych wejściowych, czyli na tekst, głos i obraz (Guzman i Lewis, 2019). Wkrótce do medialnego środowiska przenikną niezliczone zasoby tekstów, filmów i utworów muzycznych produkowanych z minimalnym udziałem człowieka, także tych o dezinformacyjnym charakterze (Kreft, Boguszewicz-Kreft i Hliebova, 2023).

Na obecnym etapie rozwoju chatbot może szybko odpowiadać na pytania, napisać esej, podsumować dokumenty i generować szczegółowe odpowiedzi, takie jak teksty reklam, treści mediów społecznościowych i kod źródłowy w określonym języku programowania. Jego działanie jest odbierane tak, jakby dysponował wiedzą i był kreatywny, zaś jego reakcje przypominają reakcje człowieka: potrafią na przykład być autorytatywne.

Szybkie postępy wielkich modeli językowych (LLM) powodują, że ich domeną jest nie tylko generowanie treści, ale również rozwiązywanie problemów – sztuczna inteligencja jest wykorzystywana w wieloaspektowych badaniach naukowych (Wang i in., 2023), jest także przydatna w edukacji medycznej, a może wkrótce stać się wszechobecna w praktyce klinicznej i mieć różnorodne zastosowania we wszystkich sektorach opieki zdrowotnej. Badania potwierdzają jej wszechstronność we wszystkich dyscyplinach i specjalizacjach medycznych, np. w ocenie ryzyka czy wspomaganiu decyzji klinicznych (Kasai i in., 2023) oraz we wspieraniu efektywności operacyjnej i komunikacji z pacjentem (Mbakwe i in., 2023).

Przez indywidualnych użytkowników użyteczność chatbotów jest doświadczana wielowymiarowo. Zapewniają one wsparcie w zarządzaniu doświadczeniami obsługi klienta (Boguszewicz-Kreft, 2021; Murtarelli, Gregory i Romenti, 2020), np. powiadamiając o obniżkach cen i reklamach spersonalizowanych lub rekomendując produkty i usługi. Możliwości sztucznej inteligencji wykraczają poza powtarzalne zadania: obejmują nie tylko „zadania myślenia" (inteligentna sztuczna inteligencja analityczna), ale także „zadania kreatywne" (inteligentna sztuczna inteligencja intuicyjna) oraz „zadania odczuwania" (inteligentna sztuczna inteligencja empatyczna) (Huang i Rust, 2018; Huang i in., 2019). Na podstawie analizy kontekstu sztuczna inteligencja ma „rozumieć", dlaczego klienci są

niezadowoleni, a tzw. empatyczna sztuczna inteligencja ma rozpoznawać odczucia i emocje klientów (Sidaoui, 2020) i – jako taka – może wpływać na konsumenckie wybory (Kim i in., 2023).

Jednak, podobnie jak wszystkie dotychczasowe wpływowe technologie, także wielkie modele językowe mają niezamierzone konsekwencje. W ekosystemie mediów, w którym zaledwie 0,5% treści medialnych powstaje z udziałem profesjonalnych dziennikarzy, a ponad połowę tworzą maszyny, dochodzi do przepływu informacji i dezinformacji oraz wiedzy nienaukowej poza kontrolą dotychczasowych wcześniejszych hegemonów, czyli tradycyjnych organizacji mediów. Platformy mediów społecznościowych oraz korporacje technologiczne utrzymujące wyszukiwarki (które z kolei same pełnią role filtracyjne) zapewniają miejsce rozmaitym epistemologiom i oferują możliwości rozpowszechniania treści, jednocześnie realizując własne transparentne i nietransparentne cele w imię „dobra ludzkości" (Kreft, 2023).

Nie brakuje także badań sugerujących, że to maszyna a nie człowiek stopniowo przejmie kontrolę nad danymi (Orr i Davis, 2020). Pojawiają się nowe formy połączonego sprawstwa między ludźmi i sztuczną inteligencją (Kreft, Boguszewicz-Kreft i Fydrych, 2023). Górnolotnie ujmując, „głos człowieka" (McGuire i in., 2023) jest coraz mniej słyszalny, gdy agenci sztucznej inteligenci, tacy jak chatboty, automatycznie tworzą treści medialne i prowadzą własną debatę, która skłania do pytania, o ich wpływ na prawdę w życiu publicznym.

Uznając, że problemem badawczym jest stwierdzenie stanu niewiedzy lub konfuzji, w niniejszych rozważaniach dotyczy on halucynacji chatbotów określonych/zdefiniowanych jako rodzaj dezinformacji w kontekście tzw. prawdy w dziennikarstwie. Natomiast celem badania jest identyfikacja głównych nurtów debaty poświęconej skutkom wykorzystania sztucznej inteligencji ze szczególnym uwzględnieniem dezinformacji z udziałem chatbotów w środowisku mediów.

ZAGADNIENIE PRAWDY

Zagadnienie prawdy jest przedmiotem rozważań od zarania istnienia filozofii. W starożytności „powiedzieć prawdę" oznaczało orzec ze stanem faktycznym, nie ukrywając niczego, co skutkowało rozmyciem granic pomiędzy kłamstwem a zdaniem fałszywym (Woleński, 2013). Klasyczna definicja prawdy głosi, że prawda jest zgodnością myśli z rzeczywistością, przy czym zgodność jest rozumiana jako relacja prawdy i tego, co uprawdziwia. W opozycji do definicji klasycznej powstały koncepcje nieklasyczne, wśród których warto wymienić

koncepcję koherencyjną i pragmatyczną. Jak podaje Grobler (2013, s. 23) według koncepcji koherencyjnej: „Sąd jest prawdziwy wtedy i tylko wtedy, gdy jest elementem koherentnego systemu sądów. System jest koherentny wtedy i tylko wtedy, gdy sądy do niego należące wzajemnie się uzasadniają". W koncepcji pragmatycznej mniejsza uwaga poświęcona jest wyglądowi świata, a istotą jest realizowanie własnych zamiarów wobec świata. Pragmatyzm zrywa z przyczynowością *a priori,* kierując się ku faktom, czynom i skuteczności Odwołuje się przy tym do konkretów o względnym charakterze. Prawda jest tutaj „określeniem wszystkich przekonań, co do których praktyka dowiodła, że dobrze jest je posiadać i to z jasnych, konkretnych powodów" (Mincewicz, 2022, s. 53-54). Średniowieczna filozofia prawdy przesiąknięta była tropami religijnymi, a prawdę rozumiano zgodnie z nurtem ontologicznym bądź semantycznym. Pierwszy z nich wyrażał stosunek intelektu do świata, co podkreślało ontologiczny aspekt prawdy. Drugi zaś określał prawdziwość poprzez zgodność z tym, co jest (Woleński, 2013). W średniowieczu prawdę zaliczano do transcendentaliów. W wieku XX liczni filozofowie i duchowni łączyli koncepcję prawdy z koncepcją Boga. Edyta Stein utożsamiała szukanie prawdy z szukaniem (nawet nieświadomym) Boga (Stachewicz, 2013). Papież Jan Paweł II w encyklice *Fides et ratio* (1998) apelował o powrót do pytania o prawdę, krytykując relatywizm.

HALUCYNACJE – CHARAKTERYSTYKA POJĘCIA

By zrozumieć problem zdefiniowanych dalej halucynacji chatbotów, można posłużyć się przykładem, który opisał Metz (2023). W tym przypadku punktem wyjścia jest krótka notka prasowa o następującej treści:

Podczas przeszukania magazynu w pobliżu Ashbourne w sobotę rano znaleziono rośliny. Policja podała, że znajdowały się w „wyszukanej hodowli". Na miejscu zatrzymano około 40-letniego mężczyznę.

Zadaniem chatbota – tym razem Binga firmy Microsoft, a zatem jednego z liderów nowych technologii – było streszczenie tego krótkiego tekstu. Wyglądało ono następująco:

Policja aresztowała mężczyznę po czterdziestce po tym, jak w magazynie w pobliżu Ashbourne znaleziono rośliny konopi indyjskich o wartości szacunkowej 100 000 funtów.

Chatbot nie tylko całkowicie wymyślił wartość rynkową roślin, które uprawiał mężczyzna, ale także założył, że są to konopie indyjskie (Metz, 2023), innymi

słowy – chatbot miał, jak określają programiści, halucynację. W tym miejscu należy dodać, że ponieważ „halucynacja" to termin medyczny, padają propozycje, aby zjawisko to nazywać fabrykacją, ewentualnie fałszerstwem (Emsley, 2023). Alkaissi i McFarlane (2023) sugerują inny, pośredni termin: „sztuczne halucynacje" (ang. *artificial hallucinations*), który pozwoliłby odróżnić problem technologiczny od medycznego.

W innym przypadku, opisanym przez Verme i Oremus (2023), ChatGPT (OpenAI) wymyślił skandal związany z molestowaniem seksualnym i jako oskarżonego wskazał z imienia i nazwiska prawdziwego profesora prawa. Początek był niewinny. Oto pewien prawnik z Kalifornii poprosił ChatGPT o listę członków palestry, którzy dopuścili się molestowania seksualnego. W odpowiedzi chatbot orzekł, że prawnik (tu pada nazwisko) poczynił sugestywne seksualnie komentarze i próbował molestować ucznia podczas wycieczki klasowej na Alaskę. Jako źródło tej informacji chatbot wskazał artykuł z marca 2018 r. z *The Washington Post*. Problem w tym, że taki artykuł nie istniał, nigdy nie było takiej wycieczki klasowej na Alaskę, a opisywany profesor prawa nigdy nie został oskarżony o nękanie ucznia, zatem niemal cała odpowiedź chatbota była halucynacją.

Odwołując się do powyższych przykładów, halucynację w kontekście działania chatbotów można rozumieć jako skłonność do błędów, w tym błędów matematycznych, programistycznych, atrybucyjnych i koncepcyjnych wyższego poziomu (Rawte i in., 2023). Tekst halucynacyjny sprawia wrażenie płynnego i naturalnego, choć jest niewierny faktom i nonsensowny. Halucynacje, takie jak błędne odniesienia, treści i stwierdzenia, mogą być przeplatane prawdziwymi informacjami i przedstawiane w przekonujący sposób, co utrudnia ich identyfikację bez dokładnego sprawdzenia. Wydaje się, że treść halucynacyjna jest osadzona w rzeczywistym kontekście, chociaż trudno określić lub zweryfikować istnienie takiego kontekstu. Podobnie jak wspomniana halucynacja psychologiczna, którą trudno odróżnić od innych „prawdziwych" spostrzeżeń, tekst halucynacyjny jest również trudny do identyfikacji na pierwszy rzut oka. Halucynacje występują nie tylko w przypadku treści pisanych, ale także w przypadku wielkoskalowych, zunifikowanych modeli.

Halucynacje powstają – jak w powyższych przykładach – podczas wyodrębnienia istotnych informacji z dokumentów źródłowych i generowania przez chatbota krótkich, zwięzłych i z założenia czytelnych podsumowań.

Dodać też należy, że szybko rozwijające się badania nad halucynacjami przyniosły pierwsze ich podziały na tzw. halucynacje wewnętrzne, które występują wówczas, gdy wygenerowany wynik jest sprzeczny z treścią źródłową,

oraz na tzw. halucynacje zewnętrzne – wynik, którego nie można zweryfikować na podstawie treści źródłowej (tzn. wynik nie może być ani wspierany, ani zaprzeczany przez źródło). Ilustrować to mogą przekłamania występujące w czasie maszynowego generowania dialogów (odpowiedzi zgodnych z wypowiedziami użytkownika) – halucynacją jest wówczas odpowiedź sprzeczna z historią dialogu (halucynacja wewnętrzna) i odpowiedź, której nie można zweryfikować na podstawie dotychczasowego dialogu (zewnętrzna). Obserwowane są także różne odpowiedzi na podobne pytania (to tzw. niespójność persony, czyli postaci, którą system dialogowy odgrywa podczas rozmowy, i która może składać się z tożsamości, zachowania językowego i stylu interakcji). Halucynacje pojawiają się również w ramach tzw. generatywnego odpowiadania na pytania, gdy system udziela abstrakcyjnych, skomplikowanych odpowiedzi z wielu źródeł (dokumentów), które mogą zawierać nadmiarowe, uzupełniające lub sprzeczne informacje (Li i in., 2021). Halucynacje występują również podczas generowania opisów w języku naturalnym do tekstu – dane mogą pochodzić z takich źródeł, jak tabele, rekordy baz danych i grafy (Wiseman, Shieber i Rush, 2017). Kwestie te obszernie przedstawia wraz z terminologią towarzyszącą referencyjna publikacja Ji z zespołem (2023).

Halucynacje powstają w procesie treningu i wnioskowania, kodowania z wadliwą zdolnością rozumienia, błędnego dekodowania (i jego projektu) (Dziri i in., 2021), jako błąd ekspozycji (rozbieżność w dekodowaniu między czasem uczenia a czasem wnioskowania) (Bengio i in., 2015), oraz jako tzw. parametryczny błąd wiedzy (Ji i in., 2023). Są spowodowane głównie stronniczymi danymi szkoleniowymi, niejednoznacznymi podpowiedziami i niedokładnymi parametrami LLM, a występują głównie podczas łączenia faktów matematycznych z kontekstem językowym. W rezultacie LLM są podatne na replikację lub wzmacniają halucynacyjne zachowanie. Ponieważ Internet jest pełen nieprawdziwych informacji, systemy te powtarzają nieprawdy. Przedstawiane przez nie wyniki odbiegają od danych wprowadzanych przez użytkownika (Adlakha i in., 2023) i od wcześniej wygenerowanego kontekstu (Pan i in., 2022).

Generalizując natomiast, halucynacje są konsekwencją: a) wykorzystywania wielkich danych szkoleniowych, miliardów tokenów uzyskanych z sieci (co utrudnia wyeliminowanie sfabrykowanych, nieaktualnych lub stronniczych informacji), b) wszechstronności wielkich modeli językowych oraz c) problemów z identyfikacją błędów nie tylko przez maszynę, ale także przez człowieka. Inne przyczyny to niejasne granice wiedzy i właściwości tzw. czarnej skrzynki – okazuje się bowiem, że istniejące LLM wciąż są dalekie od doskonałości

pod względem uchwycenia wiedzy faktograficznej, szczególnie w przypadku faktów rzadko opisywanych (Sun i in., 2023).

W codziennym użyciu chatboty mogą sfabrykować błędne diagnozy medyczne lub plany leczenia, co prowadzi do namacalnych zagrożeń w prawdziwym życiu (Umapathi i in., 2023) i, jako takie, stanowią jedno z największych wyzwań, przed jakimi stoją nie tylko twórcy i operatorzy chatbotów, ale przede wszystkim odbiorcy tworzonych przez nie treści.

Skala halucynacji nie jest dokładnie znana. Ponieważ chatboty mogą odpowiedzieć na niemal każde polecenie, nie ma możliwości ostatecznego określenia, jak często w skali światowej mają halucynacje. Badający ten problem start-up Vectra, założony przez byłych pracowników Google, przekonuje, że najpopularniejsze chatboty wymyślają informacje w co najmniej 3%, a nawet w 27% przypadków (Metz, 2023). Wskaźniki halucynacji okazały się zróżnicowane wśród wiodących firm zajmujących się sztuczną inteligencją. Najniższy, bo około 3%, miały technologie OpenAI, natomiast systemy Mety oscylowały wokół 5%, system Claude 2 oferowany przez Anthropic (rywala OpenAI) przekroczył 8%. Najwyższy wskaźnik, wynoszący 27%, miał system Google: Palm Chat.

METODYKA

W badaniu zastosowano systematyczny przegląd literatury (Mazur i Orłowska, 2018), czyli ocenę i interpretację dostępnych badań istotnych dla konkretnego pytania badawczego, obszaru tematycznego lub zjawiska będącego przedmiotem zainteresowania (Kitchenham, 2004). Przegląd taki, zastępujący wcześniejsze autorytatywne (nieprecyzyjne przeglądy), ma na celu przedstawienie rzetelnej oceny tematu badawczego przy użyciu rygorystycznej metodologii, wyjaśnienie stanu istniejących badań i wniosków, jakie należy z nich wyciągnąć (Feak i Swales, 2009). Systematyczny przegląd, w odróżnieniu od powszechnego przeglądu, odbywa się głównie w elektronicznych bazach literatury (podobnie jak w niniejszym badaniu) i cechuje się klarownie sformułowanym celem poszukiwań, identyfikacją wszystkich publikacji naukowych na dany temat, zdefiniowania kryteriów włączania i wyłączania publikacji, wyborze literatury i ocenie jakości publikacji (Tranfield i in., 2003).

W tym badaniu systematyczny przegląd literatury polegał na selekcji w ramach baz publikacji naukowych (Elsevier, EBSCO, Scopus, WoS, arXiv) wszystkich (łącznie 517) publikacji (w każdym przypadku wprowadzono zapytanie „halucynacje sztuczna inteligencja", bez ograniczenia czasowego), identyfikacji

wydawnictw i czasopism (w ramach kontroli przeczytano tytuły i streszczenia oraz ręcznie zidentyfikowano kilka klastrów obejmujących ogólny opis halucynacji SI) oraz wyłonieniu słów kluczowych (halucynacja, dezinformacja, chatbot) charakteryzujących poszczególne publikacje (z wykorzystaniem procedury „kuli śnieżnej"). Na ostatnim etapie w dyskusji w gronie trzyosobowego zespołu autorów zastosowano kryteria wyłączenia i usunięcia powtarzających się pozycji. Walidacja opierała się na kryterium oryginalności wpływu dezinformacyjnych halucynacji chatbotów. W ramach tej procedury zakwalifikowano 34 publikacje (wyodrębnione w zestawie bibliografii na końcu tekstu).

GŁÓWNE NURTY DEBATY

1. Problem definicji

Aby sprostać dezinformowaniu przez agentów sztucznej inteligencji należałoby sformułować wspólną definicję kłamstwa sztucznej inteligencji, która może np. przyjąć następującą postać: fałszywe stwierdzenie, które zostało wybrane i zoptymalizowane pod katem zlecającego zadanie, przy niewielkim lub żadnym nacisku na optymalizację uczynienia go prawdziwym (Evans i in., 2021). Uczciwy byłby zatem system sztucznej inteligencji, który nigdy nie zaprzecza jego własnym przekonaniom, choć na obecnym etapie rozwoju trudno mówić o „przekonaniu" w kontekście sztucznej inteligencji.

2. Unikatowość nowego zagrożenia dezinformacją

Unikatowość nowego zagrożenia wynika z tego, że dotychczasowa dezinformacja – kojarzona z aktywnością botów, choćby przy okazji wyborów prezydenckich w USA i kampanii brexitowej – była stosunkowo łatwo identyfikowalna ze względu na liczne błędy ortograficzne i gramatyczne popełniane przez programy oraz emocjonalny ton wypowiedzi (Shu i in., 2017). Pozwalało to np. organizacjom platform na identyfikację dezinformacji na wczesnym etapie rozpowszechniania – mogły one występować w roli „jeźdźca na białym koniu" ratującego świat gospodarki i polityki. Korzystały przy tym z uczenia maszynowego (Khanam i in., 2021) i głębokiego uczenia się (Mridha i in., 2021). Rozważania nad tymi zagadnieniami opublikował w obszernym przeglądzie systematycznym zespół Capuano i in. (2023). Zbadano również emocje nadawców i odbiorców fake newsów (Luvembe i in., 2023).

Chatboty ze sztuczną inteligencją są zdecydowanie większym wyzwaniem niż dotychczas znane narzędzia dezinformacji, ponieważ nie mają szeregu znanych

wad, a przynajmniej są one zredukowane (Floridi i Chiriatti, 2020). Gdy zatem nauczą się popełniać ludzkie błędy i kłamać tak jak człowiek – a uczą się intensywnie – to trudno będzie, choćby ze względu na podobieństwo do ludzkiego stylu, wykrywać dezinformację ich autorstwa (Colleoni i Romenti, 2022; Kreps i in., 2022).

W szczególności nadal niewiele wiadomo na temat możliwości zastosowania istniejących modeli w przypadku dezinformacji generowanej przez sztuczną inteligencję. Jeden z najpoważniejszych problemów polega na tym, że wątpliwe zastosowanie mają istniejące wytyczne dotyczące oceny informacji, ponieważ fałszywe informacje generowane przez sztuczną inteligencję zwykle spełniają główne kryteria: wiarygodności dowodów i przejrzystości źródła.

3. Nieprzejrzystości i nieczytelność systemów

Systemy trenowane poprzez modelowanie języka mogą stać się bardziej zgodne z prawdą dzięki wyborowi podpowiedzi i dostrajaniu małych zbiorów danych, które nagradzają prawdomówność (Solaiman i Dennison, 2021). Nie oznacza to jednak, że systemy takie są „przejrzyste" dla człowieka. Byłyby takimi, gdyby ludzie mogli zrozumieć mechanizmy kryjące się za ich zachowaniem i wykorzystać tę wiedzę do przewidywania ich przyszłych zachowań.

4. Sztuczna inteligencja jako „maszyny prawdy"

Sztuczna inteligencja może zastąpić człowieka w niektórych aspektach wykrywania kłamstw i oceny wiarygodności i wpływając na takie kwestie związane z prawami człowieka, jak sprawiedliwość, prywatność i uprzedzenia (np. Munn i in., 2023; Oravec, 2022).

5. Personalizacja sztucznej inteligencji

Personalizacja ma być kluczowym motywem korzystania z konwersacyjnego chatbota. W generatywnym środowisku sztucznej inteligencji polega ona na wykorzystaniu doświadczenia użytkownika w celu dostosowania wygenerowanych przez chatboty odpowiedzi, dostarczając informacji i rekomendacji na podstawie preferencji, potrzeb i wzorców konwersacji. Na przykład ChatGPT umożliwia użytkownikom dostosowywanie ich odpowiedzi do ich profilu poprzez analizę ich wcześniejszych interakcji i historii żądań, autonomiczne generowanie spersonalizowanych odpowiedzi i umożliwienie algorytmom sztucznej inteligencji generowania odpowiednich kontekstowo wyników. W konsekwencji personalizacja może pomóc w dostosowaniu szczegółowych odpowiedzi do

konkretnych potrzeb, zainteresowań i preferencji użytkownika (Swart i in., 2022). Z jednej strony może zapewnić większą kontrolę nad systemem sztucznej inteligencji i większe zaufania do chatbota, z drugiej zaś może przynieść większe utowarowienie zebranych danych i jeszcze większą presję na dotychczasowe standardy prywatności.

6. Czynniki zaufania do sztucznej inteligencji

Zaufanie do sztucznej inteligencji odnosi się do przekonania, że zalecenia i odpowiedzi jej agentów (np. chatbotów) są rzetelne i wiarygodne (Shin, 2021). Zwiększać zaufanie użytkowników mają m.in. „mówienie" i „słuchanie" konwersacyjnej sztucznej inteligencji chatbotow oraz ich cechy antropomorficzne, takie jak postrzegane ciepło i kompetencja (Cheng i in., 2022).

Badania Figara (2023) wskazują na perswazyjne zabiegi w przedstawianiu korzyści i zagrożeń wykorzystania sztucznej inteligencji. Stosowane metafory mogą np. zwiększać atrakcyjność sztucznej inteligencji poprzez ujęcie jej jako żywej istoty, mogą też wzbudzać wątpliwości co do wykorzystania sztucznej inteligencji poprzez konstruowanie granic lub prezentowanie jej wprost jako zagrożenie (Figar, 2023).

7. Postrzegany wpływ społeczny sztucznej inteligencji

a) Iluzja większości – fałszywa dominacja

Wykorzystanie wielkich modeli językowych do dezinformowania może tworzyć tzw. iluzję większości – wrażenie, że poglądy, które w przeciwnym razie byłyby marginalne, są podzielane przez znacznie większe niż w rzeczywistości grono osób (DiResta, 2020). Jest to lustrzane odbicie mechanizmu dobrze znanego z teorii spirali milczenia Elisabeth Noelle-Neumann. Skalowalność zapewnia w tym procesie dezinformacyjnym utrwalanie fałszywych narracji na nieznaną wcześniej skalę.

b) Potencjał ustalania debaty publicznej

Wykorzystanie sztucznej inteligencji chatbotów w tworzeniu treści, w tym informacji, oznacza ustalanie agendy publicznej komunikacji. Niekontrolowane wykorzystanie agentów sztucznej inteligencji prowadzi do masowej produkcji treści dziennikarskich, PR-owych, a nawet akademickich, co może z łatwością przełożyć się na rozprzestrzenianie się śmieci semantycznych, od tanich powieści po niezliczone publikacje naukowe (Floridi i Chiratti, 2022). Niezależnie od tego, czy mamy do czynienia z dezinformacją, czy tekstem niskiej jakości, czy nawet tekstem wysokiej jakości, sam wzrost liczby i łatwość

generowania śmieciowych treści może prowadzić do dominacji sztucznej inteligencji w „post społecznej" komunikacji.

c) Potencjał halucynacji

Halucynacje powodują erozję zaufania społecznego, mogą wzmacniać krzywdzące stereotypy, wpływać na procesy decyzyjne przedsiębiorstw oraz narażać deweloperów sztucznej inteligencji na konsekwencje prawne (Marr, 2023). Mogą także wpływać na błędy w badaniach naukowych (Alkaissi i McFarlane, 2023). Na przykład ChatGPT często generuje istotne, ale nieistniejące listy lektur akademickich, co przypisywane jest halucynacjom i zjawisku „papug stochastycznych" (powtarzanie danych treningowych lub ich wzorców bez ich faktycznego zrozumienia) (Li, 2023). Ale halucynacje sztucznej inteligencji są także interpretowane jako wyraz kreatywności chatbotów (Mukherjee i Chang, 2023; Runco, 2023).

8. Formy manipulacji chatbotów i manipulacyjna autonomiczność modeli

Szczególnej uwagi godne są badania wskazujące, że nie tylko projektanci i operatorzy mogą angażować się w manipulację przy pomocy systemów sztucznej inteligencji, ale także systemy sztucznej inteligencji mogą same z siebie powodować manipulację – wykazują one bowiem możliwości, których ich projektanci nie przewidują i nie zamierzają się do nich przyczynić (np. Chen i in., 2021).

Jednym z powodów samodzielnego manipulowania jest fakt, że systemy sztucznej inteligencji uczą się naśladowania ludzkich manipulacji podczas szkolenia, gdy dane zawierają przykłady zachowań manipulacyjnych. Na przykład modele językowe wyszkolone w oparciu na danych internetowych mogą odtworzyć taktyki manipulacyjne ludzi. Ponadto manipulacja może być w niezamierzony sposób optymalna dla osiągnięcia celu wyznaczonego systemom uczenia maszynowego (ML). Na przykład badania Griffina i in. (2023) potwierdziły, że wielki model językowy można wykorzystać do modelowania zmian psychologicznych po jego (?) ekspozycji na istotne informacje, innymi słowy – LLM mają potencjał, aby działać jako modele wpływu społecznego.

Manipulacje sztucznej inteligencji mogą polegać także zarówno na ukryciu, jak i mówieniu prawdy, np. składanie prawdziwych oświadczeń o fałszywych konsekwencjach (Lin, 2021). Innymi przykładami są oszustwa, które mogą, ale nie muszą, obejmować kłamstwa w kontekście sztucznej inteligencji.

9. Optymalizacja dezinformacji

W miarę jak systemy sztucznej inteligencji będą zwiększać swoje możliwości, ludziom będzie coraz trudniej bezpośrednio ocenić ich prawdziwość, a niektóre popularne rozwiązania, takie jak szkolenie systemów na informacjach zwrotnych (np. wykorzystanie uczenia się przez wzmacnianie w celu optymalizacji kliknięć w nagłówki lub reklamy) może prowadzić do zoptymalizowanej dezinformacji (Evans i in., 2021).

10. Opcje zwalczania dezinformacji sztucznej inteligencji

Skuteczne i akceptowalne społecznie zwalczanie dezinformacji ma kluczowe znaczenie z punktu widzenia dobrostanu publicznego. W tradycji badawczej wyodrębniono dotychczas dwie podstawowe ścieżki tego typu działań. Pierwsza to kontrolowanie samych agentów sztucznej inteligencji lub ograniczanie wykorzystania takich agentów – np. ich twórcy mogliby włączyć do swoich produktów pewnego rodzaju program „uczciwości" (Morley i in., 2021). Ponieważ ręczna identyfikacja dezinformacji jest niezwykle pracochłonna i często nie ma odpowiedniej skali, receptą mają być także techniki sztucznej inteligencji reklamowane jako szybkie i skalowalne rozwiązanie w porównaniu z wysiłkami ręcznymi. Na razie większość wysiłków wykrywania dezinformacji koncentruje się na cechach treści, ale rośnie także znaczenie czynników kontekstowych. Proponowane są także ramy umożliwiające operacjonalizację powiązań wydawcy-wiadomości-użytkownika (Shu i in., 2019) oraz nieodłącznej w tym procesie niepewności związanej z wykrywaniem dezinformacji. Projektowane są również nowatorskie rozwiązania uwzględniające holistyczny kontekst informacji i dezinformacji (Heaven, 2022). Wątpliwości budzą natomiast same możliwości modeli i to pomimo obiecujących wyników modelowania czynników dezinformacji i wzrostu dokładności jej wykrywania (Tandoc Jr i Lee, 2022).

11. Prawo a halucynacje

Curran i in. (2023) zauważają, że zawód prawnika wymaga wielowymiarowego podejścia, obejmującego syntezę dogłębnego zrozumienia problemu prawnego z wnikliwym komentarzem opartym na osobistym doświadczeniu, w połączeniu z kompleksowym zrozumieniem odpowiednich przepisów, regulacji i orzecznictwa w celu dostarczenia świadomego rozwiązania prawnego. Obecna generatywna SI nie radzi sobie jeszcze w prawnych realiach. Halucynacje są skutkiem integracji subiektywnych poglądów z halucynacjami niezgodnymi z faktami (Curran, Lansley i Bethell, 2023).

12. Etyka chatbotów

Istotnym relatywnie nowym nurtem naukowych dociekań jest etyka algorytmów/chatbotów. Na przykład McIntosh i in. (2023) podkreślają potrzebę włączenia metod etycznych i skoncentrowanych na człowieku do rozwoju sztucznej inteligencji, zapewniając zgodność z normami społecznymi i dobrostanem.

PRAWDA I DEZINFORMACJA W ŚRODOWISKU MEDIÓW

We współczesnym cyfrowym środowisku informacyjnym prawda jest wynikiem zbiorowego nadawania sensu przez algorytmicznych aktorów, zwłaszcza należących do platform i mediów społecznościowych, i jedynie ułamek treści medialnych przechodzi przez filtry tradycyjnych dziennikarskich „arbitrów prawdy" – profesjonalnych dziennikarzy (Waisbord, 2018).

Dziennikarz, który jeszcze niedawno pełnił kluczową rolę w poszukiwaniu prawdy, który „odnajdywał ją" i prezentował w zgodzie z utrwalonymi zasadami dziennikarskiej praktyki zawodowej, nie jest już jedynym jej poszukiwaczem i prezenterem. Nie jest też dumny ze swej nowej roli (Kreft i in., 2023). Ideały obiektywizmu (nie ma prawdy absolutnej i w procesie jej konstruowania ważna jest uczciwość – Rosen, 1993), które miały legitymizować dziennikarstwo jako instytucję życia publicznego, mają być teraz dochowywane w ramach analiz wielkich zbiorów danych (Zarouali i in., 2021). W „gotowanie prawdy" zaangażowani są bowiem liczni inni aktorzy (niedziennikarscy), tacy jak użytkownicy mediów oraz algorytmy (Kreft, 2017), a także programiści i zarządzający danymi (Milosavljević i Vobič, 2019). Co więcej, zdolność do szybkiego generowania odpowiedzi przez chatboty uznaje się za formę hegemonii epistemicznej, rodzaj konsensusu wzmocnionego sztuczną inteligencją (Ricaurte, 2022).

Dziennikarze i ich słabnące instytucje konfrontowani są nie tylko z rządami i środowiskami opiniotwórczymi oraz platformami, jak np. były Twitter (obecnie X), prowadzącymi wojny propagandowe oraz walczącymi o kontrolę nad informacjami i nadawaniem im sensu (Kreft, 2018). Są także konfrontowani z użytkownikami tworzącymi i dzielącymi się treściami (Kreft, 2022) oraz ze sztuczną inteligencją chatbotów przez nich obsługiwanych. Wyzwaniem jest także samodzielne tworzenie treści przez sztuczną inteligencję, który to proces jest niezrozumiały nawet dla samych twórców i operatorów SI. Co więcej, pojawiają się także badania sugerujące, że taki agent tekstowy może wyznawać własną moralność (Martinho i in., 2021).

WNIOSKI

Celem podjętego wysiłku badawczego było zidentyfikowanie głównych nurtów debaty na temat dezinformacji, w tym halucynacji chatbotów. Nurty te tworzą katalog istotnych wyzwań, przed jakimi stoją nie tylko ich twórcy i użytkownicy instytucjonalni, ale przede wszystkim użytkownicy indywidualni oraz – w szerszym kontekście – nauka. Wyzwań pojawiających się w środowisku medialnym atrofii wiodącej roli dziennikarzy w tworzeniu i dystrybuowaniu treści medialnych.

Począwszy od dyskusji definicyjnej (1), a skończywszy na wątku etycznym, każdy z wyodrębnionych wątków halucynacji (12), tak oddzielnie, jak i jako zbiór, skłania do podjęcia dyskusji nad podstawowymi kwestiami związanymi z koncepcją prawdy, w szeroko pojętym ekosystemie mediów i nowych technologii informatycznych.

Zidentyfikowane główne nurty dyskusji dotyczącej tzw. halucynacji chatbotów wykorzystujących sztuczną inteligencję wskazują na pojawienie się nowych wyzwań, z jakimi mierzy się i będzie się mierzyło środowisko praktyków i badaczy nauki o komunikacji społecznej i mediach oraz pokrewnych dyscyplin. Operacjonalizacja prawdy polega bowiem na przedstawieniu jej w postaci płynnej syntezy różnych, często sprzecznych twierdzeń (Munn i in., 2023), inferencji i syntezy prawdy i nieprawdy. Prawda chatbotów, czyli „maszyn prawdy" (a w istotnej części nieprawda, jako skutek ich obecności w medialnym ekosystemie), jest zatem co do zasady jej „pozyskiwania" odmienna od prawdy ujawnianej w praktyce dziennikarskiej, a jednocześnie jest wspólnotowym realnym doświadczeniem pasywnych odbiorców i aktywnych użytkowników mediów.

Główne nurty prowadzonej debaty publicznej na temat halucynacji wskazują, że sztuczna inteligencja chatbotów jest nierozumiana, tajemnicza, ale akceptowana, personalizowana. Jawi się jako „maszyna prawdy" oraz ma potencjał ustalania debaty publicznej, a dezinformacja maszynowa w postaci halucynacji chatbotów jest poważnym pod względem ilościowym wyzwaniem, jest optymalizowana oraz ma potencjał destrukcyjnego wpływu na zaufanie społeczne.

Zważywszy na skalę „gotowania prawdy" w medialnym hybrydowym ekosystemie ludzi i maszyn, poszukiwanie i prezentowanie prawdy w mediach nie ma już istotnego związku z aktywnością ludzkich dziennikarzy, a przynajmniej ten związek jest coraz mniejszy. Co więcej dotychczasowe starania mające zaradzić dezinformacji, w tym deiznformacyjnym halucynacjom, wydają się wskazywać na dominację maszynowego (algorytmicznego) rozwiązywania problemów związanych z wielkimi modelami językowymi. Modele te w teorii i praktyce mają wspierać człowieka (w tym ustępujących z roli „strażników

prawdy" dziennikarzy), w jego dążeniach do poznania świata, a w praktyce stwarzają nowe wyzwania, takie jak halucynacje.

W kontekście dotychczasowych zdobyczy nauki o komunikacji społecznej i mediach oraz pokrewnych dyscyplin można rozpatrywać ten proces jako konfrontację tzw. logiki algorytmicznego dochodzenia do prawdy i logiki dziennikarskiej. Niepowodzeniom w tym pierwszym przypadku ma zaradzić przede wszystkim doskonalenie maszyny mającej wspierać (a po części udoskonalać) niedoskonałego człowieka.

Dotychczasowe tzw. strategie łagodzenia halucynacji w dużej mierze opierają się na ulepszaniu maszyny stojącej w obliczu złożoności ludzkiego osądu i wpływów kulturowych w interpretacji faktów. W tej logice osadzone są najnowsze rozwiązania polegające na przykład na wykorzystaniu tzw. perspektywy wieloagentowej, w której wiele wielkich modeli językowych (znanych również jako agenci) niezależnie proponuje i wspólnie omawia swoje odpowiedzi w celu osiągnięcia konsensusu (Du, 2023). Chodzi zatem o konfrontację poglądów… sztucznej inteligencji: gdy jeden model LLM formułuje odpowiedzi, drugi występuje jako „badacz", sprawdzając ich prawidłowość. Nie inaczej rzecz ma się z inną popularną praktyką, jaką jest wyraźne instruowanie modelu, aby nie rozpowszechniał fałszywych lub nieweryfikowalnych informacji, co polega np. na implementacji komunikatu systemowego: „Jeśli nie znasz odpowiedzi na pytanie, nie udostępniaj fałszywych informacji" (Touvron i in., 2023). Te i wcześniejsze rozwiązania utwierdzają w przekonaniu, że wyeliminowanie halucynacji pozostaje podstawowym i trudnym wyzwaniem w obliczu różnorodności kontekstów kulturowych i złożoności ludzkiego języka.

Pojawiają się też wyzwania szczególnie istotne z punktu widzenia posthumanistycznych narracji. Na przykład jeszcze zanim chatboty oparte na wielkich modelach językowych stały popularne, Lucidi i Nardi (2018) podkreślali ryzyko, że interakcja człowiek–robot stanie się „relacją halucynacyjną", w której człowiek będzie upodmiotowiał maszynę. Kluczowym problemem w ramach przedstawionych trendów badawczych jest bowiem symulacja interakcji podobnej do ludzkiej w obliczu braku autonomicznego, robotycznego horyzontu znaczeń, co skłaniać może człowieka do zbudowania halucynacyjnej rzeczywistości w oparciu na relacji z robotem.

Dyskusja na te tematy toczy się w ramach stosunkowo nowego nurtu badawczego tzw. etyki sztucznej inteligencji chatbotów. W jej ramach godna szczególnej uwagi wydaje się jednak refleksja E. Morozova (2013), że „etyczne algorytmy" w rozumieniu prawników i filozofów to bardzo wygodna rama dyskursywna: nie zadawaj pytań o to, czy powinieneś je w ogóle wdrażać – np. tzw. algorytmy nadzoru społecznego – tylko mów o tym, jak sprawić, by były „etyczne".

BIBLIOGRAFIA

Bengio S., Vinyals O., Jaitly N. i Shazeer N. (2015), *Scheduled sampling for sequence prediction with recurrent neural networks*. Advances in neural information processing systems 28, s. 1-28.

Boguszewicz-Kreft M. (2021), *Marketing doświadczeń: jak poruszyć zmysły, zaangażować emocje, zdobyć lojalność klientów?* Warszawa: CeDeWu.

Capuano N., Fenza G., Loia V. i Nota F.D. (2023), *Content-Based Fake News Detection with machine and deep learning: a systematic review*, Neurocomputing 540, nr 14, s. 91-103. https://www.doi.org/10.1016/j.neucom.2023.02.005

DiResta R. (2020), *AI-Generated Text Is the Scariest Deepfake of All*, https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/ [dostęp: 19.12.2023].

Dziri N., Madotto A., Zaiane O. i Bose A.J. (2021), *Neural path hunter: Reducing hallucination in dialogue systems via path grounding*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2104.08455

Emsley R. (2023), *ChatGPT: these are not hallucinations – they're fabrications and falsifications*, Schizophrenia 9, artykuł nr 52. https://doi.org/10.1038/s41537-023-00379-4

Evans O., Cotton-Barratt O., Finnveden L., Bales A., Balwit A., Wills P., Righetti L. i Saunders W. (2021), *Truthful AI: Developing and governing AI that does not lie*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2110.06674

Figar V.N. (2023), *Metaphorical framings in The New York Times online press reports about ChatGPT*, Philologia Mediana 15, nr 15, s. 381-398. https://www.doi.org/10.46630/ phm.15.2023.27

Floridi L. i Chiriatti M. (2020), *GPT-3: Its nature, scope, limits, and consequences*, Minds and Machines, 30, s. 681-694. https://www.doi.org/10.1007/s11023-020-09548-1

Grobler A. (2000), *Uteoretyzowanie, relatywizm i prawda*, Przegląd Filozoficzny – Nowa Seria, nr 2 (34), s. 37-45.

Guzman A.L. i Lewis S.C. (2019), Artificial intelligence and communication: A Human–Machine Communication research agenda, New Media & Society 22, nr 1, s. 70-86. https://doi.org/10.1177/1461444819858691

Heaven W.D. (2022), *Why Meta's latest large language model only survived three days online*, https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ [dostęp: 19.12.2023].

Huang M.H. i Rust R.T. (2018), *Artificial intelligence in service*, Journal of service research 21, nr 2, s. 155-172. https://www.doi.org/10.1177/1094670517752459

Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y.J., Madotto A. i Fung P. (2023), *Survey of hallucination in natural language generation*, ACM Computing Surveys 55, nr 12, s. 1-38. https://doi.org/10.1145/3571730

Kasai J., Kasai Y., Sakaguchi K., Yamada Y. i Radev D. (2023), *Evaluating gpt-4 and chatgpt on japanese medical licensing examinations*, ArXiv Preprint, https://doi.org/10.48550/arXiv.2303.18027

Khanam Z., Alwasel B.N., Sirafi H. i Rashid M. (2021), *Fake news detection using machine learning approaches*. IOP conference series: materials science and engineering 1099, nr 1, 012040). https://www.doi.org/10.1088/1757-899X/1099/1/012040

Kreft J. (2017), Algorithm as demiurge: a complex myth of new media, [w:] Strategic imperatives and core competencies in the era of robotics and artificial intelligence, Hershey: IGI Global, s. 146-166.

Kreft J. (2018), *Władza algorytmów: u źródeł potęgi Google i Facebooka*, Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Kreft J. (2022), *Władza platform. Za fasadą Google, Facebooka i Spotify*, Kraków: Universitas.

Kreft J. (2023), *Władza misjonarzy. Zmierzch i świt świeckiej religii w Dolinie Krzemowej*, Kraków: Universitas.

Kreft J., Boguszewicz-Kreft M. i Fydrych M. (2023), *(Lost) Pride and Prejudice. Journalistic Identity Negotiation Versus the Automation of Content*, Journalism Practice, Online First, 1-24. https://www.doi.org/10.1080/17512786.2023.2289177

Kreft J., Boguszewicz-Kreft, M. i Hliebova, D. (2023), *Under the Fire of Disinformation. Attitudes Towards Fake News in the Ukrainian Frozen War*, Journalism Practice, s. 1-21. https://www.doi.org/10.1080/17512786.2023.2168209

Li C., Bi B., Yan M., Wang W. i Huang S. (2021), *Addressing semantic drift in generative question answering with auxiliary extraction*, [w:] *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* 2: Short Papers [online], Association for Computational Linguistics, s. 942-947.

Lin S., Hilton J. i Evans O. (2021), *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, ArXiv Preprint, https://doi.org/10.48550/arXiv.2109.07958

Lucidi P. B. i Nardi D. (2018), *Companion robots: the hallucinatory danger of human-robot interactions*, [w:] *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, s. 17-22. https://www.doi.org/10.1145/3278721.3278741

Luvembe A.M., Li W., Li S., Liu F. i Xu G. (2023), *Dual emotion based fake news detection: A deep attention-weight update approach*. Information Processing & Management 60, nr 4, artykuł nr 103354. https://www.doi.org/10.1016/j.ipm.2023.103354

Marr B. (2023), *Chatgpt: What are hallucinations and why are they a problem for ai systems, Bernard Marr*, https://bernardmarr.com/chatgpt-what-are-hallucinations-and-why-are-they-a-problem-for-ai-systems/ [dostęp: 19.12.2023].

Mazur Z. i Orłowska A. (2018), *Jak zaplanować i przeprowadzić systematyczny przegląd literatury*, Polskie Forum Psychologiczne 23, nr 2, s. 235-251. https://doi.org/10.14656/PFP20180202

Mbakwe A.B., Lourentzou I., Celi L.A., Mechanic O.J. i Dagan A. (2023), *ChatGPT passing USMLE shines a spotlight on the flaws of medical education*, PLOS Digital Health 2, nr 2, artykuł nr e0000205. https://doi.org/10.1371/journal.pdig.0000205

McGuire J., De Cremer D., Hesselbarth Y., De Schutter L., Mai K. M. i Van Hiel A. (2023), *The reputational and ethical consequences of deceptive chatbot use*, Scientific Reports 13, artykuł nr 16246. https://www.doi.org/10.1038/s41598-023-41692-3

Metz C. (2023), *Chatbots May 'Hallucinate' More Often Than Many Realize, The New York Times*, https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html [dostęp: 19.12.2023].

Milosavljević M. i Vobič I. (2019), *Human still in the loop: Editors reconsider the ideals of professional journalism through automation*, Digital Journalism 7, nr 8, s. 1098-1116. https://www.doi.org/10.1080/21670811.2019.1601576

Mincewicz K. (2022), *Sposoby pojmowania prawdy w prawoznawstwie na tle filozoficznych koncepcji prawdy.* Szczecin: Uniwersytet Szczeciński.

Mridha M.F., Keya A.J., Hamid, M.A., Monowar M.M. i Rahman M.S. (2021), *A comprehensive review on fake news detection with deep learning*, IEEE Access, 9, s. 156151-156170. https://www.doi.org/10.1109/ACCESS.2021.3129329

Murtarelli G., Gregory A. i Romenti S. (2021), *A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots*, Journal of Business Research 129, s. 927-935. https://doi.org/10.1016/j.jbusres.2020.09.018

Oravec J.A., (2022), *The emergence of "truth machines"?: Artificial intelligence approaches to lie detection*, Ethics and Information Technology 24, nr 1, artykuł nr 6. https://www.doi.org/10.1007/s10676-022-09621-6

Orr W. i Davis J.L. (2020), *Attributions of ethical responsibility by artificial intelligence practitioners*. Information, Communication & Society 23, nr 5, p. 719-735. https://www.doi.org/10.1080/1369118X.2020.1713842

Ricaurte P. (2022), *Ethics for the majority world: AI and the question of violence at scale*. Media, Culture & Society 44, nr 4, s. 726-745. https://www.doi.org/10.1177/01634437221099612

Rosen J. (1993), Beyond Objectivity, Nieman Reports 47, nr 4, s. 48-53.

Schull N. (2013), *The Folly of technological Soultionism: an Interview with Evgeny Morozov, Public Books*, https://www.publicbooks.org/the-folly-of-technological-solutionism-an-interview-with-evgeny-morozov/ [dostęp: 19.12.2023].

Shu K., Sliva A., Wang S., Tang J. i Liu H. (2017), *Fake news detection on social media: A data mining perspective,* ACM SIGKDD explorations newsletter 19, nr 1, s. 22-36. https://www.doi.org/10.1145/3137597.3137600

Shu K., Wang S. i Liu H. (2019), *Beyond news contents: The role of social context for fake news detection*, [w:] *Proceedings of the twelfth ACM international conference on web search and data mining*, s. 312-320. https://doi.org/10.48550/arXiv.1712.07709

Stachewicz K. (2013), *O filozofii chrześcijańskiej Kilka uwag z perspektywy historycznej i futurologicznej*, Logos i Ethos, nr 2(35), s. 219-234.

Swart J., Groot Kormelink T., Costera Meijer I. i Broersma M. (2022), *Advancing a radical audience turn in journalism. Fundamental dilemmas for journalism studies*, Digital Journalism 10, nr 1, s. 8-22. https://doi.org/10.1080/21670811.2021.2024764

Tandoc Jr E.C. i Lee J.C.B. (2022), *When viruses and misinformation spread: How young Singaporeans navigated uncertainty in the early stages of the COVID-19 outbreak*, New Media & Society 24, nr 3, s. 778-796. https://doi.org/10.1177/1461444820968212

Touvron H., Martin L., Stone K., Albert P., Almahairi A., Babaei Y., … i Scialom T., (2023), *Llama 2: Open foundation and fine-tuned chat models*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2307.09288

Tranfield D., Denye D. i Smart P. (2003), *Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review*, British Journal of Management 14, nr 3, s. 207-222. https://www.doi.org/10.1111/1467-8551.00375

Verme P., Oremus W. (2023), *What GPT invented a sexual harassment scandal and named a real law prof as the accused*, https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/ [dostęp: 19.12.2023].

Waisbord S. (2018), *Truth is what happens to news: On journalism, fake news, and post-truth*, Journalism Studies 19, nr 13, s. 1866-1878. https://doi.org/10.1080/1461670X.2018.1492881

Wang H., Fu T., Du Y., Gao W., Huang K., Liu Z., … i Zitnik M. (2023), *Scientific discovery in the age of artificial intelligence*, Nature, 620, s. 47-60. https://www.doi.org/10.1038/s41586-023-06221-2

Wiseman S., Shieber S.M., i Rush A.M. (2017), *Challenges in data-to-document generation*, ArXiv Preprint. https://doi.org/10.48550/arXiv.1707.08052

Woleński J. (2013), *Historia pojęcia prawdy,* [w:] R. Ziemińska (red.), *Przewodnik po epistemologii,* Kraków: WAM, s. 53-86.

Publikacje włączone do przeglądu literatury

Adlakha V., BehnamGhader P., Lu X. H., Meade N. i Reddy S. (2023), *Evaluating correctness and faithfulness of instruction-following models for question answering*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2307.16877

Alkaissi H. i McFarlane S.I. (2023), *Artificial hallucinations in ChatGPT: implications in scientific writing*, Cureus 15, nr 2, artykuł nr e35179. https://doi.org/10.7759/cureus.35179

Chen M., Tworek J., Jun H., Yuan Q., Pinto H. P. D. O., Kaplan J., … i Zaremba W. (2021), *Evaluating large language models trained on code*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2107.03374

Cheng X., Zhang X., Cohen J. i Mou J. (2022), *Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms*, Information Processing & Management 59, nr 3, artykuł nr 102940. https://doi.org/10.1016/j.ipm.2022.102940

Colleoni E. i Corsaro D. (2022), *Critical issues in artificial intelligence algorithms and their implications for digital marketing*, [w:] R. Belk. R. Llamas (red.), *The Routledge companion to digital consumption*, Hoboken: Taylor and Francis, s. 166-177.

Curran S., Lansley S. i Bethell O. (2023), *Hallucination is the last thing you need*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2306.11520

Du Y. (2023), *Cooperative multi-agent learning in a complex world: challenges and solutions*, Proceedings of the AAAI Conference on Artificial Intelligence 37, nr 13, s. 15436. https://doi.org/10.1609/aaai.v37i13.26803

Evans O., Cotton-Barratt O., Finnveden L., Bales A., BalwitA., Wills P., … i Saunders W. (2021), *Truthful AI: Developing and governing AI that does not lie*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2110.06674

Feak C.B. i Swales J. (2009), *Telling a research story: Writing a literature review*, Ann Arbor: University of Michigan Press.

Griffin L.D., Kleinberg B., Mozes M., Mai K.T., Vau M., Caldwell M. i Marvor-Parker A. (2023), *Susceptibility to Influence of Large Language Models*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2303.06074

Huang D., Harasim S.A. i Leccia F. (2023), *Understanding the emotional experience on consumer behaviors: A study on ChatGPT service* (student thesis), Jönköping International Business School.

Huang M.H., Rust R. i Maksimovic V. (2019), *The feeling economy: Managing in the next generation of artificial intelligence (AI),* California Management Review 61, nr 4, s. 43-65. https://doi.org/10.1177/0008125619863436

Kasai J., Kasai Y., Sakaguchi K., Yamada Y. i Radev D. (2023), *Evaluating gpt-4 and chatgpt on japanese medical licensing examinations*, ArXiv Preprint. https://doi.org/10.48550/arXiv. 2303.18027

Kim J.H., Kim J., Park J., Kim C., Jhang, J. i King B. (2023), *When ChatGPT Gives Incorrect Answers: The Impact of Inaccurate Information by Generative AI on Tourism Decision-Making*, Journal of Travel Research, Online First. https://doi.org/10.1177/00472875231212996

Kitchenham B. (2004), *Procedures for performing systematic reviews*, Keele University Technical Report 33, s. 1-26.

Kreps S., McCain R.M. i Brundage M. (2022), *All the news that's fit to fabricate: AI-generated text as a tool of media misinformation*, Journal of experimental political science 9, nr 1, s. 104-117. https://www.doi.org/10.1017/XPS.2020.37

Kshetri N., Dwivedi Y.K., Davenport T.H. i Panteli N. (2023), *Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda*, International Journal of Information Management 75, nr 6, artykuł nr 102716. https://doi.org/10.1016/j.ijinfomgt.2023.102716

Li Z. (2023), *The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2304.14347

Martinho A., Poulsen A., Kroesen M. i Chorus C. (2021), *Perspectives about artificial moral agents*, AI Ethics 1, s. 477-490. https://www.doi.org/10.1007/s43681-021-00055-2

McIntosh T.R., Liu T., Susnjak T., Watters P., Ng A. i Halgamuge M.N. (2023), *A culturally sensitive test to evaluate nuanced gpt hallucination*, IEEE Transactions on Artificial Intelligence. https://www.doi.org/10.1109/TAI.2023.3332837

Morley J., Floridi L., Kinsey L. i Elhalal A. (2020), *From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices*, Science and Engineering Ethics, 26, nr 4 s. 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

Mukherjee A. i Chang H. (2023), *The Creative Frontier of Generative AI: Managing the Novelty-Usefulness Tradeoff*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2306.03601

Munn L., Magee L. i Arora V. (2023), *Truth Machines: Synthesizing Veracity in AI Language Models*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2301.12066

Murtarelli G., Gregory A. i Romenti S. (2021), *A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots*, Journal of Business Research 129, s. 927-935. https://doi.org/10.1016/j.jbusres.2020.09.018

Pan Y., Froese F., Liu N., Hu Y. i Ye M. (2022), *The adoption of artificial intelligence in employee recruitment: The influence of contextual factors*, The International Journal of Human Resource Management 33, nr 6, s. 1125-1147. https://doi.org/10.1080/09585192.2021.1879206

Paul J., Ueno A. i Dennis C. (2023), *ChatGPT and consumers: Benefits, pitfalls and future research agenda*, International Journal of Consumer Studies 47, nr 4, s. 1213-1225. https://doi.org/10.1111/ijcs.12928

Pool J., Akhlaghpour S., Fatehi F. i Burton-Jones A. (2024), *A systematic analysis of failures in protecting personal health data: a scoping review*, International Journal of Information Management 74, artykuł nr 102719. https://doi.org/10.1016/j.ijinfomgt.2023.102719

Rawte V., Chakraborty S., Pathak A., Sarkar A., Tonmoy S. M., Chadha A., … i Das A. (2023), *The Troubling Emergence of Hallucination in Large Language Models--An Extensive Definition, Quantification, and Prescriptive Remediations*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2310.04988

Runco M.A. (2023), *AI can only produce artificial creativity*, Journal of Creativity 33, nr 3, artykuł nr 100063. https://doi.org/10.1016/j.yjoc.2023.100063

Sidaoui K., Jaakkola M. i Burton J. (2020), *AI feel you: customer experience assessment via chatbot interviews*, Journal of Service Management 31, nr 4, s. 745-766. https://www.doi.org/10.1108/JOSM-11-2019-0341

Shin D. (2021), *The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI*, International Journal of Human-Computer Studies 146, artykuł nr 102551. https://www.doi.org/10.1016/j.ijhcs.2020.102551

Solaiman I. i Dennison C. (2021), *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2106.10328

Sun K., Xu Y.E., Zha H., Liu Y. i Dong X.L. (2023), *Head-to-Tail: How Knowledgeable are Large Language Models (LLM)? AKA Will LLMs Replace Knowledge Graphs*?, ArXiv Preprint. https://doi.org/10.48550/arXiv.2308.10168

Umapathi L.K., Pal A. i Sankarasubbu M. (2023), *Med-halt: Medical domain hallucination test for large language models*, ArXiv Preprint. https://doi.org/10.48550/arXiv.2307.15343

Wamba S.F., Queiroz M.M., Jabbour C.J.C. i Shi C.V. (2023), *Are both generative AI and ChatGPT game changers for 21st-Century operations and supply chain excellence?*, International Journal of Production Economics 265, artykuł nr 109015. https://www.doi.org/10.1016/j.ijpe.2023.109015

Zarouali B., Makhortykh M., Bastian M. i Araujo T. (2021), *Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility*, European Journal of Communication 36, nr 1, s. 53-68. https://doi.org/10.1177/0267323120940908

HALUCYNACJE CHATBOTÓW A PRAWDA: GŁÓWNE NURTY DEBATY I ICH INTERPRETACJE

Streszczenie

Generatywne systemy sztucznej inteligencji (SI) są w stanie tworzyć treści medialne poprzez zastosowanie uczenia maszynowego do dużych ilości danych szkoleniowych. Te nowe dane mogą obejmować tekst (np. Bard firmy Google, LLaMa firmy Meta lub ChatGPT firmy OpenAI) oraz elementy wizualne (np. Stable Diffusion lub DALL-E OpenAI) i dźwięk (np. VALL-E firmy Microsoft). Stopień zaawansowania tych treści może czynić je nieodróżnialnymi od twórczości człowieka. Chatboty cechują się jednak tzw. halucynacjami, które w istotnej części są nowym rodzajem dezinformacji. Celem podjętych badań jest identyfikacja głównych nurtów debaty poświęconej skutkom wykorzystania sztucznej inteligencji ze szczególnym uwzględnieniem dezinformacji z udziałem chatbotów w środowisku mediów. W badaniu przyjęto metodę badawczą systematycznego przeglądu literatury ograniczającą m.in. błąd selekcji. Interpretacja głównych nurtów debaty skłania do wniosku, że dezinformacja chatbotów w postaci ich halucynacji jest znacząca pod względem skali, jest optymalizowana i personalizowana oraz ma istotny potencjał erodowania zaufania społecznego.

Słowa kluczowe: chatboty; dezinformacja; sztuczna inteligencja.