mgr Jarosław Wiśnicki
Uniwersytet Mikołaja Kopernika w Toruniu
ORCID: 0000-0002-1164-4624

https://doi.org/10.29316/9788368103243_7

BEZPIECZEŃSTWO INFORMACYJNE W KONTEKŚCIE TECHNOLOGII DEEPFAKE

INFORMATION SECURITY IN THE CONTEXT OF DEEPFAKE TECHNOLOGY

Streszczenie

Niniejsze opracowanie jest poświęcone zagrożeniom informacyjnym tworzonym w oparciu o nowe technologie komunikacyjne. Podjęta analiza podyktowana jest coraz większym wpływem procesów komunikacyjnych na globalne bezpieczeństwo. Treść opracowania koncentruje się na przestawieniu tendencji w tworzeniu zagrożeń informacyjnych będących determinantami zagrożeń społecznych. W treści rozdziału znaczące miejsce zajmuje analiza techniki nakładania na materiał źródłowy połączonych obrazów i dźwięków z wykorzystaniem metod uczenia maszynowego. Zaprezentowane zostały współczesne możliwości sztucznej inteligencji w kreowaniu opinii społecznej. Celem niniejszego opracowania jest prezentacja zagrożeń informacyjnych powodowanych rozwojem technologii widzenia komputerowego i głębokiego uczenia się.

Słowa kluczowe: bezpieczeństwo informacyjne, technologie komunikacyjne, mechanizmy wpływu społecznego, deepfake, dezinformacja

Abstract

This study is devoted to information threats arising from new communication technologies. The analysis is motivated by the growing impact of communication processes on global security. The content focuses on identifying trends in the creation of information threats that serve as determinants of broader social risks. A significant part of the chapter is devoted to the examination of techniques for superimposing combined

images and sounds onto source material using machine learning methods. The chapter also highlights the contemporary capabilities of artificial intelligence in shaping public opinion. The aim of this study is to present the information threats generated by the development of computer vision technologies and deep learning.

Keywords: information security, communication technologies, mechanisms of social influence, deepfake, disinformation

Wstęp

Podstawowym celem niniejszego opracowania jest prezentacja zagrożeń informacyjnych powodowanych rozwojem technologii widzenia komputerowego i głębokiego uczenia się. Opracowanie koncentruje się na przestawieniu roli sztucznej inteligencji w tworzeniu treści dezinformacyjnych mających wpływ na bezpieczeństwo w ujęciu narodowym, jak i międzynarodowym. Ważną część prezentowanych rozważań stanowią działania wpływu społecznego przy wykorzystaniu najnowszych zdobyczy techniki informatycznej. Autor zwraca uwagę, iż wspomniane aktywności prowadzą do zmiany postrzegania ról nadawcy i odbiorcy informacji, co determinuje głęboką transformację wykraczającą znacznie poza obszar w sensie *stricte* komunikacyjnym, a obejmuje całą sferę społeczną, polityczną, ekonomiczną a także militarną współczesnego świata. Realizacja celu podstawowego wymaga wykonania kilku zadań szczegółowych takich jak: przedstawienie procesu powstawania deepfake rozpowszechnianego w formie przekazu informacyjnego przygotowanego w sposób tendencyjny, celem wywołania określonych zachowań lub postaw, przedstawienie mechanizmów wypływu społecznego oraz wskazanie korelacji procesów informacyjnych zagrażających bezpieczeństwu komunikacyjnemu. Opracowanie oparte jest na przywołanych w dalszej części rozdziału dokumentach normatywnych oraz wnikliwej analizie dyskursu publicznego. Posiłkowano się ponadto publikacjami autora niniejszego rozdziału, dotyczącymi przedmiotowej problematyki.

Obszar badawczy podjęty w opracowaniu został ograniczony podmiotem i przedmiotem. Zakres podmiotowy stanowi technologia deepfake będąca jedną z najnowszych innowacji mających poważny wpływ na bezpieczeństwo komunikacyjne. Oznacza to, że analizie i ocenie nie będą podlegać inne technologie dezinformacyjne wykorzystujące sztuczną inteligencję do kreowania przekazu

komunikacyjnego. Podstawowe metody badań użyte w pracy to: analiza wspomnianej literatury przedmiotu, analiza i synteza logiczna oraz wnioskowanie.

Opracowanie ma charakter synkrytyczny. Z jednej strony stanowi podsumowanie i uzupełnienie dotychczasowej wiedzy teoretycznej na temat operacji wpływu społecznego[1], z drugiej – zawiera potwierdzenie owej wiedzy na podstawie analizy konkretnych działań informacyjnych z wykorzystaniem sztucznej inteligencji. Poddany analizie obszar stanowi dobre pole badawcze wpływu informacji na kreowanie opinii publicznej. Nie oznacza to, że narodziny działań informacyjnych nastąpiły w czasach nam współczesnych. Wpływ dezinformacji stosowany przez ludzi w zasadzie nie różni się od innych istot żyjących. Wprowadzanie w błąd i rozliczne mistyfikacje stosowane przez ludzi nie odbiegają od naszych odpowiedników w naturze. Wygląd, który tak pieczołowicie tworzymy, to odpowiednik barwnych piór czy falbaniastych płetw czy zjawiskowego ubarwienia. Kameleony przystosowały się, by oszukiwać, ukrywając się w gałęziach, a rzekotkom ubarwienie, utrudniające drapieżnikowi identyfikację ofiary, zapewnia przetrwanie. Z punktu podjętej analizy należy podkreślić, iż język stanowiący ówcześnie podstawę komunikacji wśród ludzi, wykształcił się stosunkowo niedawno. Nastąpiło to znacznie później od mechanizmów, które po dziś dzień wykorzystywane są z powodzeniem przez faunę czy florę. Aja Raden uważa, że istnieje hipoteza, która mówi, iż ludzie stworzyli język właśnie po to, by zwiększyć możliwości manipulacji innymi[2]. Potwierdza to fakt, iż podstęp stanowi podstawę interakcji, wrodzonej zdolności do zniekształcania rzeczywistości na własne potrzeby. Tym samym zdolność ta warunkuje zasadnicze znaczenie w procesie komunikowania.

Człowiek, jak mawiał Arystoteles, jest istotą społeczną. To właśnie dlatego, nie może on żyć poza społeczeństwem, w którym zaspokaja swoje potrzeby. Dzięki zachowaniu tej koegzystencji jedna z istotniejszych potrzeb jaką jest komunikowaniem się zostaje zaspokojona. Tym samym nasze naturalne uwarunkowania powodują nieodpartą potrzebę zaspokojona pozyskiwania

[1] Przyjmujemy tu standardowe rozumienie wpływu społecznego jako procesu, w rezultacie którego dochodzi do zmiany: zachowania, opinii lub/i emocji człowieka, wskutek tego, co robią, myślą lub czują inni ludzie (E. Aronson, T. D. Wilson, R. M. Akert, *Psychologia społeczna. Serce i umysł*, Zysk i S-ka, Poznań 1997, s. 5-6; D. Doliński, J. Maciuszek, R. Polczyk, *Wokół wpływu społecznego*, Wydawnictwo UJ, Kraków 2012.

[2] A. Raden, *Prawda o kłamstwach. Dlaczego kłamiemy, czemu wierzymy w kłamstwa, jak nas oszukują*, przeł, G. Skoczylas, Dom Wydawniczy REBIS, Poznań 2021, s. 18.

i kolportowania informacji, czyli danych prawdziwych lub opartych na zrozumieniu wpływu jakichkolwiek zniekształceń na ogólną poprawność. Jak pisze R. Kupiecki „są one częścią ludzkiego ekosystemu, nie mniej istotną od wody, powietrza i pożywienia. W podobny sposób decydują bowiem o przetrwaniu i jakości życia”[3]. Wynikiem tego coraz intensywniej tworzymy, rozpowszechniamy, a co za tym idzie zaspokajamy indywidualne i zbiorowe potrzeby komunikowania się. Pozyskiwane informacje stanowią wykładnie podejmowanych przez nas decyzji w zakresie indywidualnym, jak i zbiorowych. Tok myślowy przekonuje więc, iż dezinformując przy użyciu słów, człowiek jest w stanie wprowadzić w błąd odbiorcę fałszując rzeczywistość, kreując przyszłość a także zakłamując przeszłość. Tym samym następuje deformowanie przekazu. Jego efektywność sprowadza się do wykorzystania jednego determinantu, jakim jest wypowiedziane słowo. Jak zobaczymy dalej współcześnie, dzięki nowym technologiom komunikacyjnych, człowiek jest w stanie zmanipulować wyraz twarzy na zdjęciach i w filmach modyfikując fryzurę, ruch warg i oczu a także inne procesy właściwości fizycznych ludzkiego ciała.

Środowisko bezpieczeństwa komunikacyjnego

W warunkach dzisiejszego, globalnego przepływu informacji, nowe technologie cyfrowe odgrywają kluczową rolę w zmaganiach o ludzkie serca i umysły. Jak pisał C. Clausewitz – każda epoka ma swoje wojny. Ma je więc i nasza epoka, którą okrzyknięto epoką informacji. Dowodem może być fakt, iż na progu XXI wieku doświadczamy intensyfikacji konfliktów czwartej generacji, w których dominują operacje na płaszczyźnie informacyjnej. Tym samym wojny w noosferze stanowią główną płaszczyznę współczesnych konfliktów między zwaśnionymi stronami[4]. Na szczególna uwagę w tym obszarze zasługują determinanty jakim jest potencjał komunikacyjny oraz potencjał technologiczny. Rozwój tego ostatniego w aspekcie widzenia komputerowego i głębokiego uczenia się znacząco przyczynia się do intoksykowania środowiska informacyjnego.

---

[3] R. Kupiecki, *Dezinformacja w stosunkach międzynarodowych – geneza, cele, aktorzy, metody. Zarys problemu*, [w:] *Platforma przeciwdziałania dezinformacji – budowanie odporności społecznej badania i edukacja*, red. R. Kupiecki, T. Chłoń, F. Bryjka, K. Kozłowski, J. Misiuna, J. Podemska, P. Podemski, Dom Wydawniczy Elipsa, Warszawa 2021, s. 15-32.

[4] J. Wiśnicki, *Competition between China and the USA in the noosphere*, Reality of Politics, nr 1, 2023, s. 118.

Przestrzeń cybernetyczna w ciągu dziesięcioleci stała się czołowym medium globalnych i regionalnych operacji wpływu społecznego. Coraz więcej państw w oparciu o najnowsze zdobycze techniki potrafi zręcznie manipulować przestrzenią informacyjną, wpływając na zabarwienie przekazu i kierunek dystrybuowania go. Mechanizmy wpływu społecznego, wykorzystując sztuczną inteligencję mogą sprawiać by informacja, niczym pocisk, raziła obiekty oddziaływania stanowiąc zarzewie wzniecania nienawiści, rozpowszechniania kłamstwa, wywoływania zamieszek, kreowania konfliktów czy wzniecania wojen, a nawet niszczenia filarów życia społecznego i politycznego w państwach leżących w jej operacyjnym zainteresowaniu.

Warto podkreślić, iż upublicznienie narzędzia ChatGPT przez OpenAI w listopadzie 2022 roku zapoczątkowało rozkwit tematu sztucznej inteligencji i znacząco przyspieszyło rozwój wszelkich narzędzi opartych na algorytmach AI, w tym również otwartych i darmowych narzędzi do tworzenia deepfake'ów (i to zarówno w warstwie obrazu jak i audio). Dlatego mimo tego, że termin deepfake istnieje od dawna to rok 2023 jest w pewnym sensie przełomowym w wykorzystywaniu tej technologii. Narzędzia stały się powszechnie dostępne, a materiały deepfake w świecie TikToka i Instagramu stały się czymś zupełnie zwyczajnym.

Multidomenowość współczesnego środowiska bezpieczeństwa czyni je coraz bardziej złożonym i rozdrobnionym. Zachodzące interakcje pomiędzy obszarami funkcjonowania człowieka w ujęciu regionalnym i globalnym cechuje coraz większa niepewność i nieprzewidywalność. Tym samym w znacznej części stają się aktywnościami stosowanymi poniżej progu wojny, służąc zarówno podmiotom państwowym, jak i pozapaństwowym do osiągnięcia ich celów. Znaczenie aktywności w przedmiotowym obszarze bezpieczeństwa informacyjnego podkreśla również Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej 2020, która rozdziale dotyczącym przestrzeni informacyjnej podkreśla konieczność zbudowania na poziomie strategicznym „zdolności do ochrony przestrzeni informacyjnej (w tym do systemowego zwalczania dezinformacji), rozumianej jako przenikające się warstwy przestrzeni: wirtualnej (warstwa systemów, oprogramowania i aplikacji), fizycznej (infrastruktury i sprzętu) i poznawczej (kognitywnej)"[5].

[5] *Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej*, 2020, s. 21.

Technologia mistyfikacji

Rozwój technologii opierającej się na głębokim uczeniu maszynowym, w szczególności na modelach generatywnych, takich jak *Generative Adversarial Networks* (GANs) lub *autoregressive models* powodowany pozytywnymi zastosowaniami, takimi jak w branży filmowej do tworzenia efektów specjalnych, w medycynie by generować realistyczne obrazy medyczne do celów diagnostycznych, oraz w edukacji i rozrywce umożliwiając m.in. na ożywienie idoli z minionych czasów tworząc z nimi nowe teledyski czy sceny filmowe.

Jednakże jak każdy wynalazek narażony jest na wykorzystanie go w niekoniecznie dobrym zamyśle przyświecającym jego powstaniu. W podjętym obszarze badawczym nie odbiegło to od normy, gdyż już w 2017 roku technologia wirtualna stworzyła pierwsze fałszywe filmy pornograficzne, które zostały umieszczone na serwisie internetowym Reddit. W głównej roli, wykreowane postaci przez algorytmy uczenia maszynowego wykorzystały wizerunek min. Gal Gadot, Maisie Wiliams czy Taylor Swift. Twórca tejże mistyfikacji podpisał się pod zamieszczonym materiałem jako *deepfake*. Tym samym stało się to przyczyną używania tegoż określania, nieco później jako nazwy funkcjonującej po dzień dzisiejszy[6].

Z czasem zastosowanie przedmiotowej techniki się zwiększyło stwarzając obszar badawczy, który skutkował powstaniem ogólnej definicji, która określa deepfake jako technikę generowania fałszywych treści multimedialnych, takich jak wideo, obrazy lub dźwięki, przy użyciu zaawansowanych algorytmów uczenia maszynowego, które mogą naśladować oryginalne treści z niespotykaną wcześniej dokładnością[7]. Inne ujęcie przedstawia aspekt techniczny, który skupia się na rozwoju modeli generatywnych, takich jak GANs oraz sieci neuronowych autoregresyjnych, które są wykorzystywane do tworzenia deepfake. Badają też metody wykrywania i zwalczania deepfake[8]. W literaturze przedmiotu odnajdujemy również ujęcie tejże problematyki w naukach społecznych,

---

[6] S. Cole, *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*, portal vice.com, https://www.vice.com/en/article/reddit-fake-porn-app-daisy-ridley, [dostęp: 01.09.2023].

[7] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection*, portal arxiv.org, https://arxiv.org/abs/2001.00179, [dostęp: 01.09.2023].

[8] T. Karras, S. Laine, T. Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, portal arxiv.org, https://arxiv.org/abs/1812.04948, [dostęp: 01.09.2023].

w których to etycy i socjolodzy starają się zrozumieć, jak deepfake wpływa na społeczeństwo i jakie wyzwania etyczne stwarza. Zdaniem naukowców istotnym jest obszar wpływu na dezinformację, manipulację treściami, naruszenie prywatności i potencjalne zagrożenia dla demokracji[9]. Odniesienie do przedmiotowego terminu, znajdujemy w analizie „*Tackling deepfakes in European policy*" autorstwa Parlamentu Europejskiego, w którym to czytamy „deepfakes można najlepiej rozumieć jako podzbiór szerszej kategorii „mediów syntetycznych" generowanych przez sztuczną inteligencję, która obejmuje nie tylko wideo i audio, ale także zdjęcia i tekst"[10]. Zbieżne określenie znajdujemy w opisaniu terminu przez A. Żychowskiego, który określa deepfake jako technikę polegającą na nakładaniu na materiał źródłowy połączonych obrazów i dźwięku z wykorzystaniem metod uczenia maszynowego (głównie sieci neuronowych)[11]. Warto również przytoczyć konstatację R. Lim, która odnoszą się do zjawiska mistyfikacji z użyciem sztucznej inteligencji tak odnosi się do zjawiska: „w obszarze dezinformacji następna generacja botów będzie wyglądać i zachowywać się coraz bardziej jak prawdziwi ludzie dzięki postępom w rozpoznawaniu twarzy i przetwarzaniu języka naturalnego, które umożliwiają gromadzenie dużych ilości danych. Biorąc pod uwagę, że targetowanie indywidualne do grupy docelowej jest już dziś powszechnie praktykowane (...)"[12].

Znacznie szersze podejście prezentują Mukta i in., którzy deepfake określają jako pragmatyczne media (obrazy i filmy), które zostały zdigitalizowane w celu przedstawienia aktywności ludzkich, które w rzeczywistości nigdy nie miały miejsca. Integruje zarówno „głębokie uczenie", jak i „fałszywość". Deepfake wykorzystuje sieci neuronowe, które uczą się replikować gesty, cechy charakterystyczne, intonację mowy czy mimikę twarzy danej osoby. Dzieje się to poprzez analizę ogromnych zbiorów danych. Aby sztuczne sieci neuronów

[9] R. Chesney, D. K. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, portal paprs.ssrn.com, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954, [dostęp: 01.09.2023].

[10] Parlament Europejski, *Tackling deepfakes in European policy,* portal europarl.europa.eu, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf, [dostęp: 03.09.2023] (tłum. własne).

[11] A. Żychowski, *Metody wykrywania nagrań typu deepfake*, portal pages.mini.pw.edu.pl, https://pages.mini.pw.edu.pl/~mandziukj/2020-10-28.pdf, [dostęp: 31.08.2023].

[12] R. Lim, *Disinformation As A Global Problem – Regional Perspective*, NATO Strategic Communications Centre of Excellence, Riga 2020, s. 39.

mogły dokonać zmian w zakresie wymiany mimiki, do systemu wprowadzane są treści wideo dwóch osób. Deepfake wykorzystują system pozycjonowania twarzy i sztuczną inteligencję, aby zastąpić twarz jednej osoby wyglądem innej osoby[13].

Współcześnie tworzone aplikacje jak m.in., FaceApp, Face2Face, FaceSwap, Reface czy Doublicat umożliwiają wykorzystanie fotografii cyfrowych do modyfikowania przekazu w oparciu o technologię deepfake[14]. Są one jednak stosunkowo proste w identyfikowaniu przeprowadzonej manipulacji. To nie znaczy, że zaawansowane oprogramowanie w oparciu sztuczne sieci neuronowe np. w postaci algorytmu RefaceAI nie nastręczają problemów w odkryciu mistyfikacji.

Warto tu wspomnieć o zasobach niezbędnych do tworzenia mistyfikacji. Współcześnie twórcy deepfake'ów mogą łatwo uzyskać dostęp: do danych wrażliwych, eliminując czasochłonną pracę związaną z gromadzeniem i etykietowaniem materiałów. Co więcej, wymagana moc obliczeniowa do trenowania algorytmów uczenia maszynowego jest dostępna po niskich kosztach dzięki usługom przetwarzania w chmurze. Niektóre usługi, takie jak Google Colab, zapewniają wystarczającą moc obliczeniową do tworzenia krótkich, wysokiej jakości filmów deepfake w ciągu kilku godzin. Korzystając z wielu kont Google, można uzyskać dostęp: do znacznej ilości mocy obliczeniowej przy zerowych kosztach finansowych. Co więcej by stworzyć wysokiej jakości deepfake wystarczy więc domowy komputer lub nawet smartfon z dostępem do Internetu[15].

Dobrym przykładem skuteczności wykorzystania tej technologii w dezinformowaniu jest wideo konferencja przeprowadzona w kwietniu 2021 roku z Leonidem Wołkowem, szefem sztabu Aleksieja Nawalnego, a w zasadzie

---

[13] M. S. H. Mukta, J. Ahmad, M. A. K. Raiaan, S. Islam, S. Azam, M. E. Ali, M. Jonkman. *An Investigation of the Effectiveness of Deepfake Models and Tools*, Journal of Sensor and Actuator Networks, t. 12, nr 4, 2023, s. 61 (tłum. własne).

[14] I. Dąbrowska, *Deepfake – nowy wymiar internetowej manipulacji,* Zarządzanie Mediami, nr 2., 2020, s. 91-92; O. Wasiuta, S. Wasiuta, *FakeApp jako nowe zagrożenie bezpieczeństwa politycznego i informacyjnego*, Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate, nr 9, 2019, s. 129-130.

[15] Parlament Europejski, *Tackling deepfakes in European policy,* portal europarl.europa.eu, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf, [dostęp: 03.09.2023] (tłum. własne).

z postacią podszywającą się pod niego. Dzięki stworzonej postaci, w technologii deepfake, wprowadzono w błąd zarówno posłów izby mniejszej parlamentu Holandii, jak i polityków Wielkiej Brytanii, Litwy, Łotwy i Estonii[16]. Innym przykładem wykorzystania tej technologii jest zamieszczenie wideo wykonanego w technologii deepfake w dniu 17 marca 2022 roku. Zamieszczone nagranie, przedstawia prezydenta Ukrainy Wołodymyra Zełenskiego, który w zmanipulowanym wystąpieniu nawoływał do poddania się ludności ukraińskiej. Kolejnym przykładem jest swoista satyra, która parodiując przemówienie królowej Elżbiety II na Boże Narodzenie przedstawiona została w Channel 4[17]. Przedmiotowe nagranie stacja opublikowała w Wigilię, opatrując je komentarzem: „Czy zawsze to, co widzimy i słyszymy jest tym, czym wydaje się być?". Dyrektor Channel 4, Ian Katz, w wydanym oświadczeniu zaapelował do odbiorców zwracając uwagę, iż technologia deepfake stanowi przerażającą nową granicę w walce między dezinformacją i prawdą. Tegoroczne alternatywne życzenia, pozornie wygłoszone przez jedną z najbardziej znanych i zaufanych postaci w kraju, są mocnym przypomnieniem, że nie możemy już ufać własnym oczom[18].

Proces manipulacji

Mistyfikacje dzięki utrwalaniu danego przekazu w przestrzeni informacyjnej sprawiają, że mózg zapoznaje się ze stwierdzeniami, nawet nieprawdziwymi. Przyswaja je tak, jakby były prawdziwe. Prowadzi to do sytuacji, w której im więcej ludzi uwierzy, tym bardziej uprawdopodobni się kłamstwo, a im bardziej zostanie ono uprawdopodobnione, tym więcej osób uwierzy. Jeśli będzie ich dostatecznie dużo, powstanie swojego rodzaju „masa krytyczna" mistyfikacji, w wyniku której następuje masowa reakcja, zacierająca granice pomiędzy prawdą a fałszem. Minister propagandy Rzeszy Niemieckiej Joseph Goebbels mawiał, że „kłamstwo powiedziane tysiąc razy staje się prawdą".

[16] PAP, *Wideorozmowa ze współpracownikiem Nawalnego. Politycy z czterech krajów padli ofiarami deepfake'a*, portal tvn24.pl, https://tvn24.pl/swiat/deepfake-ktos-podszyl-sie-pod-leonida-wolkowa-wspolpracownika-aleksieja-nawalnego-5076773, [dostęp: 01.09.2023].

[17] *Deepfake Queen: 2020 Alternative Christmas Message*, portal youtube.com, https://www.youtube.com/watch?v=ivy-abd2ffm&t=22s, [dostęp: 04.09.2023].

[18] *Channel 4 pokazał fałszywe orędzie królowej Elżbiety jako ostrzeżenie przed deepfake'ami (wideo),* portal wirtualnemedia.pl, https://www.wirtualnemedia.pl/artykul/krolowa-elzbieta-oredzie-boze-narodzenie-deepfake, [dostęp: 02.09.2023].

Aja Raden tłumacząc tę zależność pisze: „jedna z heurystyk wspierających uczenie się i zapamiętywanie faktów jest efektem czystej ekspozycji, która łączy to, co znajome z prawdą”[19]. Potwierdza to fakt, iż nie musimy potwierdzać źródeł by uznać za słuszne stwierdzenie, że słońce jest większe od księżyca. Dzieje się tak w konsekwencji utrwalania, w wyniku którego nasz umysł traktuje kolejną ekspozycję intoksykowanej informacji jako potwierdzenie jej prawdziwości. Czyli: im większe rażenie informacją, nie ważne, jak absurdalną, groźną lub banalną, tym bardziej jest to przyswajane. To poczucie swojskości daje fałszywe poczucie wiarygodności przekazywanej informacji. W efekcie tych zabiegów nieprawda staje się „wiedzą powszechną”, którą powtarzają zarówno koronowane głowy, jak ich podwładni. Zależność tę nazywa się efektem iluzorycznej prawdy lub efektem powtórzenia[20], który bardzo często wykorzystywany jest nie tylko w reklamie, ale też w dyskursie publicznym.

Współcześnie powstające treści w oparciu o przedmiotową technologię stają się przyczynkiem do wieloaspektowych zagrożeń w obszarze bezpieczeństwa komunikacyjnego. Do głównych z nich należy zaliczyć dezinformowanie i manipulację informacjami poprzez tworzenie fałszywych informacji wizualnych i dźwiękowych, zwielokrotniając tym samym szerzenie fałszywych wiadomości, fikcyjnych wydarzeń czy nieprawdziwych dowodów w kwestiach społecznych, politycznych lub biznesowych. Kolejnym aspektem są szkody dla reputacji, na które są narażone zarówno osoby prywatne, jak i publiczne w sytuacji, gdy ich twarze lub głosy są umieszczane w kontekstach, które są nieodpowiednie, obraźliwe lub niezgodne z ich rzeczywistymi przekonaniami. Zniesławiające treści mogą być wykorzystywane do nękania i zastraszania, podważania zaufania, co może spowodować poważne szkody reputacyjne i psychologiczne. Innym aspektem są oszustwa i przestępstwa finansowe, które dokonywane mogą być w wyniku użycia technologii fałszującej komunikację audio lub wideo. Główne zagrożenie w tym względzie sprowadza się do utraty danych osobowych, tożsamości, środków finansowych lub utraty marki czy reputacji. Następnym aspektem jest naruszenie prywatności, poprzez tworzenie fałszywych materiałów, które wydają się przedstawiać prawdziwe sytuacje lub

[19] A. Raden, *Prawda o kłamstwach. Dlaczego kłamiemy, czemu wierzymy w kłamstwa, jak nas oszukują*, przeł. G. Skoczylas, Dom Wydawniczy REBIS, Poznań 2021, s. 230.

[20] B. Czekała, *Iluzja wiedzy – skąd się bierze i jak jej zapobiec?*, portal universeofmemory.pl, https://universeofmemory.pl/iluzja-wiedzy, [dostęp: 02.09.2023].

zachowania. Jeszcze innym aspektem jest podważanie wiarygodności źródeł informacji mogące przyczynić się do spadku zaufania do źródeł informacji i mediów, powodowane powątpiewaniem w autentyczność komunikowanych treści. Skutkować to może utratą stabilności gospodarczej, powątpiewaniem w słuszność decyzji sędziowskich czy rekomendacji naukowych. Istotnym są również zagrożenia w obszarze demokracji, powodowane wprowadzeniem niejednoznaczności przekazu komunikacyjnego w procesach politycznych i wyborczych poprzez tworzenie fałszywych wystąpień publicznych lub wypowiedzi polityków. Przykład takiego zniekształcenia informacji znajdujemy w nagraniu przedstawiającym oświadczenie Jarosława Kaczyńskiego, w którym polityk pozornie popiera protesty przeciwko wyrokowi Trybunału Sprawiedliwości i krytykuje Kościół katolicki[21]. Choć nie jest klasyczny przykład techniki deepfake, bo jego powstanie oparte jest na technologii Morph Cut wykorzystującej sztuczną inteligencję w celu generowania obrazu pomiędzy klatkami, co sprawia wrażenie upłynnienia materiału. Przytoczony przykład potwierdza coraz częstsze wytwarzanie dezinformacji przy użyciu AI.

Warto podkreślić również, iż kolportujące informacje dziennikarstwo poprzez wprowadzenie „sztucznego" obrazu może doprowadzić do zacierania się granic między rzeczywistością a fikcją, utrudniając tym samym weryfikację rzetelności przekazywanych wiadomości i podważeniem wiarygodności mediów. W tym miejscu, wydaje się niezbędne odniesienie się do zagrożenia, które skutkować może powyższymi determinantami objawiającymi ogólnym odczuciem dezorientacji i niestabilności prowadzącą do erozji zaufania publicznego. Bez wątpienia utrata bezpieczeństwa komunikacyjnego w tym obszarze może doprowadzić do destabilizacji społecznej.

Korespondujące stanowisko odnajdujemy w dokumencie NATO „Strategic Communications Centre of Excellence is multi-nationally constituted and NATO-accredited international military organization", który określa powstający trend dezinformacyjny w następujący sposób: „wśród wielu obszarów budzących obawy związane z tą technologią, być może jednym z najczęściej

[21] *Jarosław Kaczyński nie poparł Strajku Kobiet. Jak rozpoznać deepfake?,* portal demagog.org.pl, https://demagog.org.pl/analizy_i_raporty/jaroslaw-kaczynski-nie-poparl-strajku-kobiet-jak-rozpoznac-deep-fake, [dostęp: 18.10.2023].

omawianych jest zagrożenie stwarzane przez deepfake: syntetyczny dźwięk, obrazy i wideo generowane przy użyciu sztucznych inteligencja”[22].

Przestrzeń cybernetyczna staje się coraz istotniejszym obszarem komunikacyjnym w którym są dystrybuowane informacje oraz zachodzą w niej interakcje społeczne. Treści, które są szeroko udostępniane, komentowane i „polubiane”, znacząco wpływają na sposób myślenia kreując tym samym poglądy społeczeństw. Platformy te umożliwiają również formowanie społeczności o podobnych interesach i przekonaniach, co może prowadzić do tzw. filtrów bańkowych i wzmacniania istniejących przekonań. Aktywności takie wykorzystywane są nader często do wpływać na opinię publiczną, kształtować narrację czy dezinformując obiekty informacyjnego oddziaływania. Ponadto istotnym w podjętym rozważaniu jest fakt, iż mimo zwiększania liczebności narzędzi do rozpoznawania deepfake’ów wzrasta liczba narzędzi do ich generowania. Powoduje to coraz większe utrudnienia w odróżnieniu materiału „sztucznego” od prawdziwego wytworzonego poprzez „zwykłego śmiertelnika”. W efekcie czego powstaje zupełnie nawa broń, która znajduje coraz szersze zastosowanie w wojnach w noosferze.

Podsumowanie

Innowacyjność technologii, ich wyjątkowość, postępowość, niezbędność oraz nasycenie środowiska informacyjnego kreatorami i rezonatorami komunikacyjnymi stwarza podatny grunt dla powstawania i rozprzestrzeniania się deepfake’ów. Stanowią one specyficzny typ oszustwa wykorzystujący sztuczną inteligencję do wprowadzania w błąd złośliwie albo w celu uzyskania korzyści materialnych lub sytuacyjnych. W tym miejscu wydaje się niezbędne odniesienie się zagrożenia bezpieczeństwa, które na skutek uzyskania przez pozapaństwowy podmiot kontroli, nawet częściowej, w przestrzeni informacyjnej państwa. Konsekwencją przejęcia inicjatywy może być intoksykowanie środowiska informacyjnego mającego wpływ na postrzeganie przez nie otoczenia i w efekcie na kształtowanie podejmowanych decyzji. Z tak dookreślonych argumentów aspekt informacyjny urasta do rangi kluczowego składnika polityki bezpieczeństwa informacyjnego państwa. W przypadku braku wystarczającej tarczy opiniotwórczej społeczeństwa, każde państwo jest narażone na konsekwencje

[22] T. Hwang, *Deepfakes – Primer and Forecast*, NATO Strategic Communications Centre of Excellence, Riga 2020, s. 4.

nie mniej kosztowne niż długotrwałe oddziaływanie gospodarcze, polityczny przymus czy konflikt zbrojny. To przekonanie wzmacnia nieustanne nasycenie środowiska informacyjnego przekazem medialnym, będącym w dzisiejszych czasach kreatorem opinii publicznej. Tym samym konieczne jest, co dostrzegają wspomniane raporty, Unii Europejskiej i Paktu Północnoatlantyckiego, zwiększenie aktywności w aspekcie bezpieczeństwa komunikacyjnego.

Jednakże do obrony przed deepfake nie wystarczy ich dementowanie, konieczna jest rozwijanie technologii ich wykrywania, które mogą pomóc w identyfikowaniu fałszywych treści. Istotną jest również edukacja i podnoszenie świadomości w zakresie istnienia technologii mogących być wykorzystane do deformowania przekazu. Podnoszenie świadomości, poprzez specyfikację zagrożeń i sposobów rozpoznawania podejrzanych treści może pomóc ludziom unikać pułapek związanych z dezinformacją. W szeroko pojętej tarczy antydezinformacyjnej społeczeństwa warto rozważyć również, z jednej strony zachęcanie do weryfikacji źródeł informacji i upewniania się, że informacje są wiarygodne, z drugiej do rozwijania technologii autentyczności. Nie oznacza to rezygnacji z kontroli nad własnymi danymi wrażliwymi, które mogą być wykorzystane do tworzenia deepfake. Zapewne znacznym wsparciem byłoby również udoskonalanie algorytmów mediów społecznościowych w zakresie rozwijania i wdrażania technologii umożliwiających automatyczne wykrywanie i oznaczanie treści deepfake, a także ograniczanie ich rozpowszechniania. Jednakże same rozwiązania technologiczne są niewystarczające. Wydaje się zasadnym podjęcie prac nad regulacjami prawnymi związanymi z wykorzystaniem deepfake w celach szkodliwych lub wyprowadzających w błąd odbiorców.

Autor wyraża nadzieję, iż przedmiotowe opracowanie stanie się przyczynkiem do naukowej refleksji i będzie źródłem twórczej inspiracji do dalszych badań. Pamiętać przy tym należy, iż badania wiążą się z pogłębianiem własnej wiedzy o badanym przedmiocie oraz propagowaniem wyników badań nie tylko w wąskim kręgu naukowym, lecz również wśród szerszego audytorium. Wiedza może być bowiem także orężem, szczególnie w aspekcie bezpieczeństwa komunikacyjnego w dobie prowadzonych wojen hybrydowych.

Literatura

1. Aronson, E., Wilson T. D., R. M. Akert, *Psychologia społeczna. Serce i umysł*, Zysk i S-ka, Poznań 1997.
2. Doliński D., Maciuszek J., Polczyk R., *Wokół wpływu społecznego*, Wydawnictwo UJ, Kraków 2012.
3. Hwang T., *Deepfakes – Primer and Forecast*, NATO Strategic Communications Centre of Excellence, Riga 2020.
4. Kupiecki R. *Dezinformacja w stosunkach międzynarodowych – geneza, cele, aktorzy, metody. Zarys problemu*, [w:] *Platforma przeciwdziałania dezinformacji – budowanie odporności społecznej badania i edukacja*, red. R. Kupiecki, T. Chłoń, F. Bryjka, K. Kozłowski, J. Misiuna, J. Podemska, P. Podemski, Dom Wydawniczy Elipsa, Warszawa 2021.
5. Lim R., *Disinformation As A Global Problem – Regional Perspective,* NATO Strategic Communications Centre of Excellence, Riga 2020.
6. Mukta M. S. H., Ahmad J., Raiaan M. A. K., Islam S., Azam S., Ali M. E., Jonkman M. *An Investigation of the Effectiveness of Deepfake Models and Tools*. Journal of Sensor and Actuator Networks, t. 12, nr 4, 2023.
7. Raden A., *Prawda o kłamstwach. Dlaczego kłamiemy, czemu wierzymy w kłamstwa, jak nas oszukują*, przeł. G. Skoczylas, Dom Wydawniczy REBIS, Poznań 2021.
8. *Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej*, Warszawa 2020.
9. Wiśnicki J., *Competition between China and the USA in the noosphere*, Reality of Politics, nr 1, 2023.
10. Wiśnicki J., *Czwarta władza w kontekście konfliktów zbrojnych i aktów terrorystycznych*, Wydawnictwo Adam Marszałek, Toruń 2019.

Netografia

1. *Channel 4 pokazał fałszywe orędzie królowej Elżbiety jako ostrzeżenie przed deepfake'ami (wideo),* portal wirtualnemedia.pl, https://www.wirtualnemedia.pl/artykul/krolowa-elzbieta-oredzie-boze-narodzenie-deepfake, [dostęp: 02.09.2023].
2. Chesney R., Citron D. K., *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, portal paprs.ssrn.com, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954, [dostęp: 01.09.2023].

3. Cole S., *We are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*, portal vice.com, https://www.vice.com/en/article/reddit-fake-porn-app-daisy-ridley, [dostęp: 01.09.2023].
4. Czekała B., *Iluzja wiedzy – skąd się bierze i jak jej zapobiec?,* portal universeofmemory.pl, https://universeofmemory.pl/iluzja-wiedzy, [dostęp: 02.09.2023].
5. *Deepfake Queen: 2020 Alternative Christmas Message*, portal youtube.com, https://www.youtube.com/watch?v=ivy-abd2ffm&t=22s, [dostęp: 04.09.2023].
6. *Jarosław Kaczyński nie poparł Strajku Kobiet. Jak rozpoznać deepfake?,* portal demagog.org.pl, https://demagog.org.pl/analizy_i_raporty/jaroslaw-kaczynski-nie-poparl-strajku-kobiet-jak-rozpoznac-deep-fake, [dostęp: 18.10.2023].
7. Karras T., Laine S., Aila T. A., *A Style-Based Generator Architecture for Generative Adversarial Networks,* portal arxiv.org, https://arxiv.org/abs/1812.04948, [dostęp: 01.09.2023].
8. PAP, *Wideorozmowa ze współpracownikiem Nawalnego. Politycy z czterech krajów padli ofiarami deepfake'a,* portal tvn24.pl, https://tvn24.pl/swiat/deepfake-ktos-podszyl-sie-pod-leonida-wolkowa-wspolpracownika-aleksieja-nawalnego-5076773, [dostęp: 01.09.2023].
9. Tolosana R., Vera-Rodriguez R., Fierrez J., A. Morales, J. Ortega-Garcia, *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection*, portal arxiv.org, https://arxiv.org/abs/2001.00179, [dostęp: 01.09.2023].
10. Żychowski A., *Metody wykrywania nagrań typu deepfake*, portal pages.mini.pw.edu.pl, https://pages.mini.pw.edu.pl/~mandziukj/2020-10-28.pdf, [dostęp: 31.08.2023].