Roczniki Kulturoznawcze
Tom XVI, numer 4 – 2025
https://doi.org/10.18290/rkult25164.16

MONIKA NOWIKOWSKA

DEEPFAKE W NOWYCH MEDIACH – ZAGROŻENIE DLA TOŻSAMOŚCI CYFROWEJ

1. *DEEPFAKE* – ISTOTA ZJAWISKA

Pojęcie *deepfake* wywodzi się z języka angielskiego jako połączenie dwóch słów: *deep learning* oraz *fake*. Słowa te w połączeniu oznaczają ‘zastosowanie metody uczenia maszynowego’ i uzyskanie ‘fałszywego wyniku’. Powszechnie *deepfake* kojarzony jest z produktem cyfrowym (audio, wideo lub nieruchomym obrazem), który został stworzony za pomocą metod uczenia maszynowego i ma charakter sztucznego wytworu.

Technologia *deepfake* zyskała popularność w 2017 roku, kiedy to użytkownik o pseudonimie *DeepFake* opublikował na forum dyskusyjnym Reddit filmy pornograficzne z celebrytami, w tym aktorką Gal Gadot i piosenkarką Taylor Swift[1]. Okazało się, że jego podstęp polegał na wklejeniu twarzy znanych osób w miejsce twarzy aktorów istniejącego filmu pornograficznego. Mimo szybkiej reakcji środowiska, film osiągnął dużą popularność i stał się pewnego rodzaju trendem. Przyjmuje się, że to zdarzenie otworzyło przed użytkownikami internetu nowe narzędzie do zabawy i manipulacji na globalną skalę[2]. Od twórcy tego zdarzenia pochodzi używane dzisiaj powszechnie pojęcie *deepfake*[3].

Dr Monika Nowikowska – Akademia Sztuki Wojennej; adres do korespondencji: al. gen. A. Chruściela „Montera” 103, 00-910 Warszawa; e-mail: m.nowikowska@akademia.mil.pl; ORCID: https://orcid.org/0000-0001-5166-8375.

[1] Ilona Dąbrowska, „Deepfake – nowy wymiar internetowej manipulacji”, *Zarządzanie mediami* 8, nr 2 (2020):91.

[2] Piotr Wasilewski i Weronika Lenart, „Technologia deepfake wymaga regulacji – komputerowe przeróbki stają się zagrożeniem”, dostęp 14.12.2024, https://www.prawo.pl/biznes/pierwsze-regulacje-prawne-technologii-deepfake,497454.html.

[3] Ilona Dąbrowska, „Deepfake – nowy wymiar internetowej manipulacji”, *Zarządzanie mediami* 8, nr 2 (2020):91.

Artykuły w czasopiśmie dostępne są na licencji Creative Commons Uznanie autorstwa – Użycie niekomercyjne – Bez utworów zależnych 4.0 Międzynarodowe (CC BY-NC-ND 4.0)

Należy podkreślić, że dostęp do technologii *deepfake* staje się coraz powszechniejszy i łatwiejszy w użyciu. Darmowe oprogramowanie i popularne aplikacje na smartfony, takie jak FaceSwap lub Zao, pozwalają zwykłym użytkownikom na tworzenie i dystrybucję tego rodzaju treści. Szybki rozwój technologii oraz łatwy dostęp do narzędzi spowodowały, że w sieci pojawiło się wiele spreparowanych filmów z udziałem polityków czy osób powszechnie znanych. *Deepfaki* szybko stały się powszechnie rozpoznawalną koncepcją wśród ogółu internautów i częstym tematem w mediach społecznościowych. W literaturze przedmiotu wskazuje się, że technologia *deepfake* została „zdemokratyzowana"[4].

Pojęcie *deepfake* zostało uregulowane na gruncie Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji)[5]. Zgodnie z treścią art. 3 pkt 60 Rozporządzenia 2024/1689 „deepfake" oznacza wygenerowane przez AI lub zmanipulowane przez AI obrazy, treści dźwiękowe lub treści wideo przypominające istniejące osoby, przedmioty, miejsca, podmioty lub zdarzenia, które odbiorca mógłby niesłusznie uznać za autentyczne lub prawdziwe.

W motywie 134 aktu w sprawie sztucznej inteligencji wskazano, że podmioty, które wykorzystują system AI do generowania obrazów, treści dźwiękowych lub wideo albo manipulowania nimi, tak by łudząco przypominały istniejące osoby, przedmioty, miejsca, podmioty lub wydarzenia i które to treści mogą niesłusznie zostać uznane przez odbiorcę za autentyczne lub prawdziwe, powinny również jasno i wyraźnie ujawnić – poprzez odpowiednie oznakowanie wyniku, że źródłem tych treści jest AI, że te treści zostały sztucznie wygenerowane lub zmanipulowane. Prawodawca unijny wyraźnie wskazał, że spełnienie obowiązku w zakresie przejrzystości nie powinno być interpretowane jako wskazujące na to, że wykorzystanie systemu AI lub jego wyników ogranicza prawo do wolności wypowiedzi czy prawo do wolności sztuki i nauki zagwarantowane w Karcie, w szczególności w przypadku, gdy te treści stanowią część dzieła lub programu mającego wyraźnie charakter twórczy, satyryczny, artystyczny, fikcyjny lub analogiczny, z zastrzeżeniem odpowiednich zabezpieczeń w zakresie praw i wolności osób trzecich. W motywie wskazano, że określony w rozporządzeniu obowiązek w zakresie przejrzystości dotyczący treści typu *deepfake* ogranicza się do ujawniania informacji o istnieniu takich

[4] Hannah Smith i Katherine Mansted, ed., *Weaponised Deep Fakes: National Security and Democracy* (Australia: Australian Strategic Policy Institute, 2020), 3-18.

[5] Dz.Urz.UE L. z 12.7.2024.

wygenerowanych lub zmanipulowanych treści w odpowiedni sposób, który nie utrudnia wyświetlania utworu lub korzystania z niego, w tym jego normalnego wykorzystania i użytkowania, przy jednoczesnym zachowaniu użyteczności i jakości utworu.

Zgodnie z dyspozycją art. 50 ust. 4 aktu w sprawie sztucznej inteligencji, podmioty stosujące system AI, który generuje obrazy, treści dźwiękowe lub wideo stanowiące treści *deepfake* lub który manipuluje takimi obrazami lub treściami, są zobowiązane do ujawnienia, że treści te zostały sztucznie wygenerowane lub zmanipulowane. Na mocy tego artykułu prawodawca nakłada obowiązek informowania o pochodzeniu i odpowiedniego oznakowania wytworów *deepfake*. W przypadku, gdy treść stanowi część pracy lub programu o wyraźnie artystycznym, twórczym, satyrycznym, fikcyjnym lub analogicznym charakterze, obowiązek w zakresie przejrzystości ogranicza się do ujawnienia istnienia takich wygenerowanych lub zmanipulowanych treści w odpowiedni sposób, który nie utrudnia wyświetlania lub korzystania z utworu. Ten obowiązek nie ma zastosowania w przypadku, gdy wykorzystanie jest dozwolone na mocy prawa w celu wykrywania przestępstw, zapobiegania im, prowadzenia postępowań przygotowawczych w ich sprawie lub ścigania ich sprawców.

Zjawisko *deepfake* stanowi także przedmiot analiz przedstawicieli polskiej i zagranicznej doktryny. Na problem prawny omawianego zjawiska zwrócono uwagę w Raporcie „Bezpieczeństwo narodowe i demokracja", opracowanym przez Australijski Instytut Polityki Strategicznej w 2020 roku. Wskazano w nim, że *deepfake* to cyfrowe fałszerstwo stworzone za pomocą sztucznej inteligencji. *Deepfake* może tworzyć całkowicie nowe treści lub manipulować istniejącymi, w tym wideo, obrazami, dźwiękiem i tekstem[6]. Jest to zagadnienie, które wymaga odpowiednich regulacji prawnych. Nobert Young definiuje *deepfake* jako technologię wykorzystującą sztuczną inteligencję do tworzenia lub edytowania treści wideo albo obrazu w celu pokazania czegoś, co nigdy się nie wydarzyło[7]. Podobnie Jon Bateman wskazuje, że *deepfake* to media generowane przez sztuczną inteligencję, które przedstawiają wymyślone wydarzenia, czasami całkiem realistycznie[8]. Na uwagę zasługuje także definicja Mette-Marie Zacher Sørensen, która zauważa, że *deepfake* to fałszywy film (syntetyczna audiowizualna produkcja medialna)

[6] Smith, Mansted, *Weaponised Deep Fakes: National Security and Democracy*, 5.

[7] Nobert Young, *DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media* (New York: Independently published, 2019), 14.

[8] Jon Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios* (Washington: Carnegie Endowment for International Peace 2020), 4.

wyprodukowany przy użyciu technik głębokiego uczenia (sztucznych sieci neuronowych)[9].

Na gruncie doktryny polskiej wskazuje się, że *deepfake* to zmanipulowany materiał audiowizualny. Ta technologia opiera się na algorytmie, który najpierw uczy się tzw. cech dystynktywnych danej osoby (jej mimiki, barwy głosu), a następnie synchronizuje je z dowolnie wybranym przez twórcę materiałem tak, aby stworzyć iluzję wskazującą, że osoba przedstawiona na filmie faktycznie wypowiada słowa stanowiące podkład dźwiękowy[10]. Piotr Wasilewski i Weronika Lenart wskazują, że

> *deepfake* to przeróbka wideo lub nagrania głosowego, która wykorzystuje sztuczną inteligencję do kreacji czegoś, co nie miało miejsca w rzeczywistości. Za pomocą zbioru danych konkretnej osoby, takich jak jedno zdjęcie i kilkusekundowe uchwycenie głosu, możemy stworzyć zupełnie nowe i całkowicie fikcyjne zachowanie danej osoby[11].

Podsumowując powyższe rozważania, można stwierdzić, że *deepfake*, jako termin pochodzący ze slangu, nie ma uzgodnionej definicji technicznej[12]. Najczęściej odnosi się do sfabrykowanego filmu lub dźwięku przedstawiającego osobę mówiącą lub robiącą coś, czego nigdy nie powiedziała[13].

Analiza przedmiotowych definicji pozwala na wskazanie dwóch elementów charakteryzujących *deepfake*: procesu głębokiego uczenia się oraz fałszerstwa.

Deepfake to technika syntezy obrazów oparta na sztucznej inteligencji. *Deepfake* jest tworzony przez zespół algorytmów głębokiego uczenia, znany jako GAN (*Generative Adversarial Network*)[14]. System ten składa się z dwóch sztucznych sieci: generatora i dyskryminatora. Sieci te pozostają w relacji konkurencyjnej. Generator tworzy fałszywe dane, dyskryminator natomiast ocenia, czy i na ile wynik jest wiarygodny. Innymi słowy, GAN to system uczenia maszynowego, który sam określa, czy wygenerowane przez niego fałszywe dane wyjściowe są wystarczająco realistyczne, a jeśli nie, dokonuje ich dalszej optymalizacji[15].

---

[9] Mette-Marie Zacher Sørensen, „Deepfake Face-Swap Animations and Affect", w: *Human Perception and Digital Information Technologies: Animation, the Body, and Affect*, red. Tamari Tomoko (Bristol: Bristol Univeristy Press, 2024), 195.

[10] Katarzyna Chałubińska-Jentkiewicz i Monika Nowikowska, *Prawo mediów* (Warszawa: Wydawnictwo C.H. Beck, 2022), 126.

[11] Wasilewski, Lenart, „Technologia deepfake wymaga regulacji – komputerowe przeróbki stają się zagrożeniem".

[12] Bateman, *Deepfakes and Synthetic Media*, 4.

[13] Bateman, *Deepfakes and Synthetic Media*, 4.

[14] Smith, Mansted, *Weaponised Deep Fakes: National Security and Democracy*, 5.

[15] Smith, Mansted, *Weaponised Deep Fakes: National Security and Democracy*, 8.

Sieci konkurują ze sobą aż do momentu, w którym nie można już odróżnić prawdziwych wyników od fałszywych. W przypadku, gdy celem jest naśladowanie powszechnie znanej osoby, w pierwszej kolejności gromadzone są dane na jej temat, które służą do trenowania sieci neuronowej.

System obliczeniowy zwany głęboką siecią neuronową pozyskuje dane treningowe (próbki) twarzy lub głosu osoby docelowej, a następnie stosuje algorytmy w celu wyodrębnienia wzorców matematycznych z danych. Opierając się na tych wzorcach, sieć generuje nowe, syntetyczne reprezentacje twarzy lub głosu. Sieć neuronowa jest zatem w stanie zidentyfikować specyficzne cechy danej osoby, które następnie mogą być wykorzystane do stworzenia unikalnego nagrania obrazu, dźwięku lub wideo tej osoby. Zebrane informacje są następnie wykorzystywane do stworzenia nowych, sztucznych treści, które odpowiadają zamierzonemu celowi, ale w rzeczywistości nie są oparte na oryginalnym obrazie lub nagraniu.

Deepfake jest technologią umożliwiającą tworzenie fałszerstw. Umożliwia stworzenie realistycznych fałszerstw przedstawiających wypowiedzi i działania, które nigdy nie miały miejsca w rzeczywistości. Fałszywe informacje generowane przy użyciu *deepfaków* mogą być obecnie wykorzystywane do wywierania wpływu na proces wyborczy, zniekształcania przekazu medialnego, napięcia społeczne, czyny nieuczciwej konkurencji, a także do innych działań dezinformacyjnych zakłócających normalne funkcjonowanie państwa i jednostek. W ten sposób powstaje zupełnie nowy poziom zagrożeń związanych z rozpowszechnianiem nieprawdziwych informacji[16].

Jako realny przykład zagrożenia można wskazać możliwość wygenerowania przez sieci GAN wiarygodnych imitacji głosu. Podszycie pod daną osobę stwarza ryzyko, że fałszywy aktor, z powodzeniem naśladujący dyrektora generalnego, jest w stanie oszukać pracowników firmy. Takie samo zdarzenie może wystąpić w jednostkach wojskowych, gdzie wykorzystany zostanie fałszywy głos dowódcy.

Do cech zjawiska *deepfake* można zaliczyć szybkość, z jaką ta technologia staje się bardziej wyrafinowana i szeroko dostępna. Początkowo treści tworzone przez sztuczną inteligencję były niedoskonałe i miały liczne błędy. Odbiorcy mieli świadomość nieautentyczności przedstawianego obrazu. Te wady mogły wynikać z niewystarczającej liczby danych dostarczonych przez ludzi algorytmom sztucznej inteligencji, co mogło prowadzić do łączenia niekompatybilnych obrazów. W efekcie w początkowych fazach rozwoju *deepfaków* zauważalne były przypadki rozmytych oczu lub niedopasowanych okularów. Należy podkreślić, że ciągły rozwój technologii prowadzi do doskonalenia analizowanego

[16] Chałubińska-Jentkiewicz, Nowikowska, *Prawo mediów*, 127.

zjawiska. Dążenie do autentyczności obrazu stawia twórcom wyzwanie doskonalenia algorytmów sztucznej inteligencji i danych wejściowych, w celu stworzenia takiego wizerunku, aby ocena jego prawdziwości nie była wykrywalna dla ludzi bez użycia specjalistycznego sprzętu.

Należy podkreślić, że technologia *deepfake* staje się coraz powszechniej dostępna dla ogółu społeczeństwa. Dedykowane programy i zestawy twarzy do pobrania powodują, że coraz większa liczba użytkowników internetu może samodzielnie tworzyć fałszywe treści. Tym samym przekonanie o niezwykłości analizowanego zjawiska ulega osłabieniu.

Wreszcie, analiza opublikowanych *deepfaków* pozwala stwierdzić, że początkowo były one tworzone dla żartu, w celu satyrycznym. Inne przedstawiały osoby publiczne w negatywnym świetle, naruszając dobra osobiste osób na nich przedstawianych, a jeszcze inne stworzono, aby poprzeć konkretny punkt widzenia[17].

W tym miejscu można postawić następujące pytania: czy zjawisko *deepfake* może stanowić poważne zagrożenie w cyberprzestrzeni? Czy obrazy kreowane przez GAN mogą być wykorzystywane do działań przestępczych? Czy i kiedy *deepfake* może stanowić zagrożenie dla tożsamości cyfrowej? Dalsze rozważania należy rozpocząć od udzielenia odpowiedzi na postawione pytania.

2. *DEEPFAKE* JAKO ZAGROŻENIE DLA TOŻSAMOŚCI CYFROWEJ

Udzielając odpowiedzi na pierwsze pytanie (czy *deepfake* może stanowić poważne zagrożenie w cyberprzestrzeni), należy podkreślić, że to zjawisko w nowych mediach dynamicznie się rozwija. Na problematyczny charakter przedmiotowego zjawiska zwróciła uwagę Organizacja Traktatu Północnoatlantyckiego (NATO). W opublikowanym 8 listopada 2019 roku przez Centrum Doskonałości Komunikacji Strategicznej NATO raporcie zatytułowanym „*The Role of deepfakes in mailing influence campaigns*”[18] podkreślono wszechobecność technologii, ale także wskazano szereg zagrożeń związanych z jej rozpowszechnianiem. Na podstawie znanych przykładów, w tym *deepfaku* z użyciem wizerunku Baracka Obamy, wskazano, że zademonstrowana zdolność do stworzenia wiarygodnej imitacji osoby publicznej przy użyciu łatwo dostępnych danych powinna wywoływać poważne zaniepokojenie potencjalnym wykorzystaniem *deepfaków* do np. manipulacji politycznej czy też dezinformacji.

[17] Dąbrowska, „Deepfake – nowy wymiar”, 91.

[18] Keir Giles, Kim Hartmann i Munira Mustaffa, *The Role of deepfakes in mailing influence campaigns* (Riga: NATO Strategic Communications Centre of Excellence, 2019), 8.

Przedmiotowy raport omawia zjawisko *deepfake* jako fenomen współczesnych technologii. Autorzy, dostrzegając popularność nowej technologii, podkreślają podstawowe zagrożenie z nią związane, a mianowicie oszustwo. Korzystanie z tej technologii przykładowo może mieć wpływ na demokratyczne wybory. W przypadku, gdy oponent polityczny za pomocą *deepfaków* przygotuje dystrybucję krótkiego filmu rozpowszechniającego „fałszywkę" i nie będzie wystarczająco dużo czasu, aby druga strona mogła rzetelnie się do niej odnieść, może to wpływać na decyzje wyborcze obywateli. *Deepfaki* mogą być wykorzystywane do dostarczania fałszywych, ale wiarygodnych wiadomości, tak jakby pochodziły od znanych osób. Zatem cyfrowa manipulacja wideo może sprawić, że fałszywe wiadomości, kreowane na prawdziwe, mogą mieć realny wpływ na podejmowane przez obywateli decyzje, w różnych aspektach ich życia[19].

W literaturze przedmiotu wskazuje się, że celem dezinformacji jest przekazanie fałszywej informacji odbiorcy, który będzie przekonany, że jest ona prawdziwa. Przy zastosowaniu technologii *deepfake*, przy jednoczesnym braku możliwości odniesienia się przez odbiorcę do oryginałów, bardzo trudne lub niekiedy niemożliwe staje się zweryfikowanie autentyczności informacji lub obrazu[20].

Zawarte w analizowanym raporcie wyniki badań mogą również pomóc w udzieleniu odpowiedzi na drugie pytanie, a mianowicie, czy obrazy kreowane przez GAN mogą być wykorzystywane do działań przestępczych. W Raporcie opisano kazus Katie Jones – fałszywej tożsamości utworzonej na profilu LinkedIn z użyciem technologii *deepfake*. W 2019 roku podająca się za młodą badaczkę z Waszyngtonu Katie Jones założyła profil na LinkedIn, wypełniając dane biograficzne, takie jak posiadane stopnie naukowe czy zajmowane stanowisko w prestiżowym think tanku. W profilu zamieszczono także zdjęcie. Było całkowicie wygenerowanym komputerowo artefaktem wykonanym przy użyciu algorytmów uczenia maszynowego. Posiadając tak zbudowaną tożsamość, Katie Jones rozpoczęła budowanie sieci kontaktów.

W Raporcie wskazano, że 52 osoby przyjęły zaproszenia od Katie. Nawiązała ona kontakt z przedstawicielami różnych think tanków, naukowcami, oficerami wojskowymi i urzędnikami państwowymi. Wśród najwyższych rangą

[19] Krzysztof Marek Kiełpiński, „Deepfake jako narzędzie do przekazywania informacji fałszywej i domniemanej. Analiza prawnokarna i cybernetyczna", *Kwartalnik Krajowej Szkoły Sądownictwa i Prokuratur*y 3, nr 51 (2023):85; Olga Wasiuta i Sergiusz Wasiuta, „Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość", *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate* 9 (2019):19; Olga Wasiuta i Sergiusz Wasiuta, „FakeApp jako nowe zagrożenie bezpieczeństwa politycznego i informacyjnego", *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate* 9 (2019):129.

[20] Kiełpiński, „Deepfake jako narzędzie do przekazywania informacji fałszywej", 85.

profesjonalistów znalazły się takie osoby, jak były generał jednogwiazdkowy i attaché obrony USA w Moskwie, najwyższy rangą urzędnik Departamentu Stanu USA, dyrektor amerykańskiej firmy chłodniczej z siedzibą w Moskwie. Profil założony w marcu 2019 roku został zidentyfikowany jako fałszywy na początku kwietnia i publicznie ujawniony w czerwcu. W raporcie podkreślono, że nieznany jest cel, w jakim została stworzona fałszywa tożsamość. Wskazano równocześnie, że fałszywe konto zaprzestało swojej aktywności, zanim stało się obiektem podejrzeń, co może wskazywać, że Katie zrealizowała zamierzony cel, zanim została wykryta.

W wyniku analizy przedmiotowego kazusu, jako przykładowe cele, dla których mogła zostać stworzona fałszywa tożsamość Katie Jones, wskazano zaawansowany *spearphishing*, uzyskanie internetowego lub fizycznego dostępu do wydarzeń, mapowanie sieci, działanie testowe lub żart.

Spearphishing to bardziej wysublimowana forma phishingu[21]. Przestępcy przed jego przeprowadzeniem wykonują wnikliwą pracę wywiadowczą, aby uzyskać jak najwięcej informacji o osobie lub grupie osób będących celem oszustwa dla zwiększenia skuteczności swoich działań. Przy tej metodzie działania oszuści podszywają się pod konkretne osoby lub organizacje, które ofiara zna i którym ufa, wysyłają do niej fałszywe wiadomości, często zawierające informacje z życia prywatnego, w celu zwiększenia ich wiarygodności. W przeciwieństwie do tradycyjnej formy phishingu, *spearphishing* jest zdecydowanie bardziej skuteczny. Wymaga od oszustów większego zaangażowania, a także niemal bezpośredniego kontaktu z ofiarą. Powszechnie wskazuje się, że dobrze przygotowany atak może zmylić nawet najbardziej świadomego zagrożeń użytkownika sieci[22]. Fikcyjny profil Katie, wykreowany na platformie LinkedIn, miał charakter profesjonalnie przygotowanej tożsamości, mającej wzbudzać zaufanie. Takie działanie mogło mieć na celu wykreowanie zaufanego źródła do korespondencji z wybranymi

[21] Zob. Monika Nowikowska, „Digital Identity on the Internet – Challenges and Threats", w: *Wielowymiarowość Cyberbezpieczeństwa*, red. Justyna Żylińska, Katarzyna Huczek i Krzysztof Borkowski (Warszawa: Uczelnia Techniczno-Handlowa im. Heleny Chodkowskiej, 2024), 28; Monika Nowikowska, „Identity Theft. Protection of Personal Data in Cyberspace", w: *Digital well -being − a concern for the quality of life*, red. Laura Tafaro, Ildiko, Laki i Iwona Florek (Józefów: Wyższa Szkoła Gospodarki Euroregionalnej im. Alcide De Gasperi w Józefowie and Milton Friedman University, 2023), 148; Filip Radoniewicz, „Phishing", w: *Leksykon cyberbezpieczeństwa*, red. Katarzyna Chałubińska-Jentkiewicz (Warszawa: Wydawnictwo Akademii Sztuki Wojennej, 2024), 198; Filip Radoniewicz, *Odpowiedzialność karna za hacking i inne przestępstwa przeciwko danym komputerowym i systemom informatycznym* (Warszawa: Wydawnictwo C.H. Beck, 2016), 108.

[22] Mariusz Nowak, „Spearphishing – czym różni się od standardowego phishingu?", dostęp 04.03.2025, https://www.netia.pl/pl/srednie-i-duze-firmy/youtro-strefa-wiedzy/spear-phishing-co-to-jest.

osobami, w celu dostarczenia im przykładowo złośliwego lub szpiegującego oprogramowania na ich urządzenia za pośrednictwem poczty elektronicznej lub innego systemu przesyłania wiadomości. W przypadku Katie Jones, będącej całkowicie sztuczną postacią, wyzwanie było prostsze. Treść musiała być po prostu realistyczna i wiarygodna, a nie jedynie podobna do konkretnej osoby.

Jako drugi cel, który mógł przyświecać stworzeniu fikcyjnej tożsamości, wskazano możliwość uzyskania internetowego lub fizycznego dostępu do wydarzeń, informacji, danych, poprzez zapisanie się na listy i otrzymywanie w ten sposób istotnych powiadomień oraz poświadczeń.

Mapowanie sieci stanowiło kolejny potencjalny cel wykreowania fikcyjnej tożsamości. Mapowanie sieci polega na gromadzeniu informacji o powiązaniach między osobami w określonych obszarach badań nad polityką.

Wykreowanie fałszywej tożsamości Katie Jones mogło mieć także znaczenie testowe. Uruchomienie testowe polega na ocenie wiarygodności tego rodzaju profilu, sprawdzeniu jego skuteczności w penetracji i budowaniu sieci oraz kontaktów, jak również testowaniu jego wykrywalności w celu ukierunkowania na prowadzenie podobnych kampanii w przyszłości. To działanie mogło pokazać, jakie profile mają szanse na sukces oraz wykorzystanie ich do manipulacji[23].

Wreszcie, jako piąty cel stworzenia fałszywej tożsamości wskazano zwykły żart. Takie działanie mogło być podjęte dla rozrywki twórcy fikcyjnej tożsamości.

Udzielając odpowiedzi na ostatnie pytanie (czy i kiedy *deepfake* może stanowić zagrożenie dla tożsamości cyfrowej jednostki), należy wskazać, że już pierwsze wykorzystanie filmu z użyciem technologii *deepfake* w 2017 roku stanowiło naruszenie wizerunku Gal Gadot oraz Taylor Swift. Ta technologia w swej istocie zakłada posłużenie się wizerunkiem lub głosem danej osoby, tworząc iluzję, że dana osoba mówi lub robi coś, czego w rzeczywistości nie zrobiła. Zatem technologia *deepfake* z natury rzeczy stanowi poważne zagrożenie dla tożsamości jednostki.

Należy jednak podkreślić, że nie każde użycie technologii *deepfake* będzie stanowiło naruszenie tożsamości cyfrowej jednostki. Od 2017 roku, tj. od powstania zjawiska, można wskazać na kilka obszarów przestrzeni cyfrowej, w której ta technologia jest wykorzystywana. *Deepfake* był wykorzystywany w produkcjach kinematograficznych, m.in. takich, jak „Gra o tron" czy „Gwiezdne wojny". W tych przypadkach *deepfake* wykorzystywany był do stworzenia cyfrowej repliki. Ta technologia pozwalała na to, aby nieżyjącego aktora umieścić w kolejnym wątku serii, która cieszyła się popularnością wśród widzów[24].

---

[23] Wasilewski i Lenart, „Technologia deepfake wymaga regulacji".

[24] Kiełpiński, „Deepfake jako narzędzie do przekazywania informacji fałszywej", 90.

Omawiana technologia będzie miała charakter przestępczy, kiedy jej wykorzystanie ma na celu niszczenie reputacji konkretnej osoby oraz naruszenie jej dóbr osobistych. W literaturze przedmiotu wyróżnia się cztery typy *deepfaków*: rozrywkowe, edukacyjne, dezinformacyjne oraz dyskredytacyjne. *Deepfake* w obszarze rozrywki ma na celu rozbawienie odbiorcy, na wzór satyry lub karykatury. W obszarze edukacji technologia *deepfake* może wykorzystywać wizerunek osób historycznych w celu przybliżenia odbiorcom faktów z ich życia. Trzeci obszar jest związany z dezinformacją. Użycie *deepfake* ma wywołać szum medialny oraz niepokój społeczny. W tym kontekście przede wszystkim wykorzystuje się wizerunki osób publicznych w celu wywołania określonych zachowań. Jako przykład można wskazać zastosowanie technologii *deepfake* w 2022 roku na początku konfliktu pomiędzy Federacją Rosyjską a Ukrainą. 16 marca tego roku za pomocą technologii *deepfake* zaatakowano kanał telewizyjny Ukraine 24, na którym transmitowano wystąpienie prezydenta Ukrainy Wołodymyra Zełenskiego. Podczas przemówienia rzekomo wzywał swoich rodaków do porzucenia broni w obliczu rosyjskiej agresji. W analizowanym przypadku przy użyciu technologii *deepfake* przekazano nieprawdziwą informację. Ostatni typ ma charakter dyskredytacyjny. Wykorzystuje się go w przestrzeni publicznej w celu naruszenia dobrego imienia osoby, której wizerunek został wykorzystany.

Zjawisko *deepfake* stanowi jedno z narzędzi, które może być wykorzystywane do naruszenia tożsamości w sieci. Wydaje się, że w miarę jego rozwoju należy podejmować praktyczne kroki o charakterze łagodzącym i regulacyjnym. Ustawodawstwo krajowe powinno określać, czy i kiedy oszustwo dokonywane za pomocą *deepfaków* jest lub powinno być przestępstwem, a jeśli tak, to jakim. Również platformy mediów społecznościowych powinny mieć obowiązek prawny reagowania i być wzywane do zajęcia się niektórymi z najbardziej szkodliwych konsekwencji działalności prowadzonej w ich sieciach. Jednak najpotężniejsza obrona przed możliwym szkodliwym wpływem *deepfaków* pozostaje taka sama, jak przed złośliwymi kampaniami wpływu: świadomość oraz odpowiednio rozwinięte i dobrze poinformowane postrzeganie zagrożeń. Istotna jest tu edukacja każdej jednostki. Po stronie państwa leży obowiązek organizowania kampanii uświadamiających w zakresie cyberbezpieczeństwa, edukacji ogółu społeczeństwa, która powinna obejmować przystępne wyjaśnienia natury i konsekwencji technologii *deepfake*. Jako przykład kampanii edukacyjnych można wyprodukować własne demonstracyjne filmy typu *deepfake*, publikowane w kontrolowanych okolicznościach, ilustrujące ich potencjał do oszukiwania w celu edukowania swoich odbiorców. Wreszcie każda osoba fizyczna powinna być świadoma, że jej wizerunek lub informacje opublikowane publicznie

w internecie mogą stać się treścią cyfrową możliwą do wykorzystania w niegodziwych celach[25].

3. WNIOSKI

Wyzwania, przed jakimi stoi ochrona tożsamości w sieci, nie jest sytuacją statyczną, ale rozwijającym się i dynamicznym procesem. Podejście cyberprzestępców stale ewoluuje i rozwija się. Wynika z tego, że te państwa, które przygotowują się do przeciwdziałania obecnie widocznym zagrożeniom, mogą wkrótce uznać swoje strategie za nieaktualne. Reakcje na wszystkie przypadki naruszania tożsamości jednostki w sieci muszą być przemyślane, ciągłe, wyczulone na trendy i ukierunkowane na przyszłość, aby przeciwdziałać potencjalnym przyszłym zagrożeniom. Na podstawie przeprowadzonych rozważań można stwierdzić, że w najbliższej przyszłości należy spodziewać się kilku kluczowych trendów. Po pierwsze, szybki rozwój i wdrażanie tzw. kampanii wpływu wykorzystujących nowe technologie. Kampanie wpływu odgrywają coraz większą rolę w kształtowaniu decyzji obywateli dotyczących ważnych aspektów ich życia. Po drugie, dalszy rozwój algorytmów uczenia maszynowego uwiarygadniających profile i interakcje w celu budowania sieci w celach komercyjnych lub złośliwych. Po trzecie, brak jasnych i skutecznych regulacji w zakresie nowych zjawisk, takich jak przykładowo *deepfake*, może skutkować większym wykorzystywaniem do celów politycznych czy działań nielegalnych. Po czwarte, równolegle, obok prób regulacji w celu przeciwdziałania zjawiskom niepożądanym, postępować będzie rozpowszechnianie technologii *deepfake* jako powszechnie akceptowanej, w obszarze rozpowszechnienia osób wirtualnych, zwłaszcza w marketingu i reklamie. Wreszcie, ostatnim trendem może być brak zaufania społecznego dotyczący tego, czy interakcje online są rzeczywiście prowadzone z prawdziwą osobą.

Wydaje się, że wyścig między tworzeniem a wykrywaniem *deepfaków* będzie trwał, a każda ze stron będzie cieszyć się chwilową przewagą. Pozorny sukces w rozwoju technik wykrywania może dawać fałszywą pewność, że problem został rozwiązany. Jak trafnie wskazują autorzy Raportu NATO Strategic Communications Centre of Excellence, K. Giles, K. Hartmann, M. Mustaffa, „*deepfaki* są jak terroryzm”[26]. Niemożliwe już jest ich całkowite wyeliminowanie. Ważne jest znalezienie sposobów na życie z nimi jako odwiecznym problemem i złagodzenie najbardziej szkodliwych prawdopodobnych skutków. Podobnie jak w walce

[25] Giles, Hartmann i Mustaffa, *The Role of deepfakes in mailing influence campaigns*, 23.

[26] Giles, Hartmann i Mustaffa, *The Role of deepfakes in mailing influence campaigns*, 25 (tł. własne).

z terroryzmem, twórcy *deepfaków* będą szukać nowych sposobów zaskoczenia. Wprowadzane ograniczenia prawne mogą być niewystarczające do ograniczania elastyczności i pomysłowości w opracowywaniu nowych sposobów wykorzystania technologii do wyrządzania szkód. W związku z tym *deepfaki* mogą z czasem stanowić kluczowy element wykorzystania strategii informacyjnych w cyberprzestrzeni w przyszłych wojnach i konfliktach hybrydowych.

BIBLIOGRAFIA

Bateman, Jon. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington: Carnegie Endowment for International Peace, 2020.

Chałubińska-Jentkiewicz, Katarzyna i Monika Nowikowska. *Prawo mediów*. Warszawa: Wydawnictwo C.H. Beck, 2022.

Chałubińska-Jentkiewicz, Katarzyna i Monika Nowikowska. „Informacja w mediach. Dezinformacja, trolling, postprawda, fake news, deepfake". *Edukacja Prawnicza* 2, nr 181 (2022):16-24.

Dąbrowska, Ilona. „Deepfake – nowy wymiar internetowej manipulacji". *Zarządzanie mediami* 8, nr 2 (2020):89-101.

Giles Keir, Kim Hartmann i Munira Mustaffa. *The Role of deepfakes in mailing influence campaigns*. Riga: NATO Strategic Communications Centre of Excellence, 2019.

Kiełpiński, Krzysztof Marek. „Deepfake jako narzędzie do przekazywania informacji fałszywej i domniemanej. Analiza prawnokarna i cybernetyczna". *Kwartalnik Krajowej Szkoły Sądownictwa i Prokuratur*y 3, nr 51 (2023):83-99.

Nowak, Mariusz. „Spear phishing – czym różni się od standardowego phishingu?". Dostęp 04.03.2025. https://www.netia.pl/pl/srednie-i-duze-firmy/youtro-strefa-wiedzy/spear-phishing-co-to-jest.

Nowikowska, Monika. „Digital Identity on the Internet – Challenges and Threats". W: *Wielowymiarowość Cyberbezpieczeństwa*, red. Justyna Żylińska, Katarzyna Huczek, Krzysztof Borkowski, 25-38. Warszawa: Uczelnia Techniczno-Handlowa im. Heleny Chodkowskiej, 2024.

Nowikowska, Monika. „Identity Theft. Protection of Personal Data in Cyberspace". W: *Digital well-being − a concern for the quality of life*, red. Laura Tafaro, Ildiko Laki, Iwona Florek, 148-165. Józefów: Wyższa Szkoła Gospodarki Euroregionalnej im. Alcide De Gasperi w Józefowie and Milton Friedman University, 2023.

Radoniewicz, Filip. „Phishing". W: *Leksykon cyberbezpieczeństwa*, red. Katarzyna Chałubińska-Jentkiewicz, 198. Warszawa: Wydawnictwo Akademii Sztuki Wojennej, 2024.

Radoniewicz, Filip. *Odpowiedzialność karna za hacking i inne przestępstwa przeciwko danym komputerowym i systemom informatycznym*. Warszawa: Wydawnictwo C.H. Beck, 2016.

Smith, Hannah i Katherine Mansted. Red. *Weaponised Deep Fakes: National Security and Democracy*. Australia: Australian Strategic Policy Institute, 2020.

Sørensen, Mette-Marie Zacher. „Deepfake Face-Swap Animations and Affect". W: *Human Perception and Digital Information Technologies: Animation, the Body, and Affect*, red. Tamari Tomoko, 195-212. Bristol: Bristol Univeristy Press, 2024.

Wasilewski, Piotr i Weronika Lenart. „Technologia deepfake wymaga regulacji – komputerowe przeróbki stają się zagrożeniem". Dostęp 14.12.2024. https://www.prawo.pl/biznes/pierwsze-regulacje-prawne-technologii-deepfake,497454.html.

Wasiuta, Olga i Sergiusz Wasiuta. „Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość". *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate* 9 (2019):19-30.

Wasiuta, Olga i Sergiusz Wasiuta. „FakeApp jako nowe zagrożenie bezpieczeństwa politycznego i informacyjnego". *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate* 9 (2019):129-139.

Young, Nobert. *DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media*. New York: Independently published, 2019.

DEEPFAKE W NOWYCH MEDIACH – ZAGROŻENIE DLA TOŻSAMOŚCI CYFROWEJ

Streszczenie

Nasze możliwości technologicznie stale idą na przód, nie zawsze jednak mogą stanowić powód do radości. Obok wielu korzyści wynikających z tego faktu, trzeba być także świadomym nowych zagrożeń. W tym przypadku można mówić o zjawisku *deepfake*. Jeszcze do niedawna szczytem dezinformacji było rozpowszechnianie fałszywych zdjęć i tekstów. Wraz z rozwojem ery cyfrowej, rozwinęły się także możliwości sztucznej inteligencji, która osiągnęła zupełnie nowy poziom i jest obecnie w stanie wykreować także sztuczne wideo. Od 2017 roku zjawisko *deepfake* pojawiło się w przestrzeni wirtualnej. Należy podkreślić, że w literaturze przedmiotu brakuje monografii, która kompleksowo opisywałyby zagrożenia związane z technologią *deepfake*. Artykuł opisuje zjawisko *deepfake* jako narzędzie do przekazywania nieprawdziwej informacji oraz zagrożenie dla tożsamości cyfrowej. Ta technologia w swej istocie zakłada posłużenie się wizerunkiem lub głosem danej osoby, tworząc iluzję, że dana osoba mówi lub robi coś, czego w rzeczywistości nie zrobiła. Zatem technologia *deepfake* z natury rzeczy stanowi poważne zagrożenie dla tożsamości jednostki. W tym celu użyto kilku metod badawczych, w tym: dogmatycznoprawnej, historycznej i porównawczej. Wśród wniosków najważniejszy jest ten, że *deepfake* jest nową formą manipulacji i dezinformacji. W przyszłości *deepfaki* mogą stanowić kluczowy element wykorzystania strategii informacyjnych w cyberprzestrzeni. Niemożliwe już jest ich całkowite wyeliminowanie. Ważne jest znalezienie sposobów na życie z nimi jako odwiecznym problemem i złagodzenie najbardziej szkodliwych prawdopodobnych skutków.

Słowa kluczowe: deepfake; dezinformacja; nowe technologie; tożsamość cyfrowa; wojna informacyjna

DEEPFAKES IN NEW MEDIA – A THREAT TO DIGITAL IDENTITY

Summary

Our technological capabilities are constantly advancing, but this is not always a reason to celebrate. Alongside the many benefits that come with this, we also need to be aware of new threats. In this case, we are talking about the phenomenon of deepfake. Until recently, the dissemination of fake photos and texts was the pinnacle of disinformation. With the development of the digital age, the possibilities of artificial intelligence have also developed to a whole new level and are now able to create artificial videos. Since 2017, the phenomenon of deepfake has appeared in virtual space. It should be emphasized that there is no monograph in the literature that comprehensively describes the dangers associated with deepfake technology. The article describes the phenomenon of deepfake as a tool for communicating false information and a threat to digital identity. This technology essentially involves using a person's image or voice to create the illusion that the person is saying or doing something that they have not actually done. Deepfake technology therefore inherently poses a serious threat to an individual's identity. Several research methods were used for this purpose, including dogmatic-legal, historical and comparative. Among the conclusions, the most important is that deepfake is a new form of manipulation and disinformation. In the future, deepfakes may be a key element in the use of information strategies in cyberspace. It is no longer possible to eliminate them completely. It is important to find ways to live with them as an eternal problem and to mitigate the most harmful probable effects.

Keywords: deepfake; disinformation; new technologies; digital identity; information warfare