Cybersecurity and Law

2025 Nr 2 (14)

DOI: 10.34567/cal/215986

Deepfake technology: definition, characteristics of the phenomenon, and mechanisms of operation

Technologia Deepfake: podstawy definicyjne, charakterystyka zjawiska oraz mechanizmy funkcjonowania

Grzegorz PILARSKI
Akademia Sztuki Wojennej
ORCID: 0000-0001-9728-2611
E-mail: g.pilarski@akademia.mil.pl

Monika NOWIKOWSKA
Akademia Sztuki Wojennej
ORCID: 0000-0001-5166-8375
E-mail: m.nowikowska@akademia.mil.pl

Piotr GALEMBA
Akademia Sztuki Wojennej
ORCID: 0000-0003-0551-4185
E-mail:p.galemba@doktorant.akademia.mil.pl.

Abstract

The aim of this article is to analyse deepfake technology, with particular emphasis on the characteristics of the phenomenon in question, to define it, describe how it works, compare it with other forms of digital manipulation, and indicate its significance in the context of social threats. The work focuses on identifying the features that distinguish deepfake from other forms of digital manipulation and discussing the factors influencing its growing effectiveness and prevalence.

The research method used was an analysis and critical review of the literature on the subject, including scientific publications, institutional reports and expert studies on generative technologies and digital disinformation. This method was supplemented by a comparative analysis of selected definitions and classifications of deepfakes, which allowed for a synthetic approach to the phenomenon and its placement in a broader social and technological context.

The results of the analyses indicate that deepfakes represent a qualitatively new stage in the development of digital manipulation, characterised by a high level of realism, automation of the creation process and the ability to imitate individual human characteristics such as appearance, voice and facial expressions. The mechanisms of

this technology are based on dynamic learning of artificial intelligence models on large data sets, which leads to a gradual blurring of the line between authentic and synthetic content. At the same time, it has been shown that the pace of development of generative tools currently exceeds the capabilities of their effective detection.

In conclusion, it was accepted that deepfake technology poses a significant challenge to contemporary information societies, affecting the credibility of media communication, social trust and information security. An interdisciplinary approach is therefore necessary, combining the development of manipulation detection technologies, appropriate legal regulations and media education, in order to limit the negative consequences of this phenomenon while preserving its potential for use.

Keywords

deepfake, disinformation, manipulation, artificial intelligence, image

Streszczenie

Celem niniejszego artykułu jest analiza technologii deepfake ze szczególnym uwzględnieniem charakterystyki przedmiotowego zjawiska, próby określenia definicji, mechanizmów funkcjonowania, oraz wskazania znaczenia w kontekście zagrożeń społecznych. Praca koncentruje się na identyfikacji cech odróżniających deepfake od innych form manipulacji cyfrowej oraz na omówieniu czynników wpływających na jego rosnącą skuteczność i powszechność.

Zastosowaną metodą badawczą była analiza i krytyczny przegląd literatury przedmiotu, obejmujący publikacje naukowe, raporty instytucjonalne oraz opracowania eksperckie dotyczące technologii generatywnych i dezinformacji. Metodę tę uzupełniono analizą porównawczą wybranych definicji i klasyfikacji deepfake'ów, co pozwoliło na syntetyczne ujęcie zjawiska oraz jego osadzenie w szerszym kontekście społecznym i technologicznym.

Rezultaty przeprowadzonych analiz wskazują, że deepfake stanowi jakościowo nowy etap rozwoju manipulacji cyfrowej, charakteryzujący się wysokim poziomem realizmu, automatyzacją procesu tworzenia oraz zdolnością do naśladowania indywidualnych cech ludzkich, takich jak wygląd, głos czy mimika. Mechanizmy funkcjonowania tej technologii opierają się na dynamicznym uczeniu modeli sztucznej inteligencji na dużych zbiorach danych, co prowadzi do stopniowego zacierania granicy pomiędzy treściami autentycznymi a syntetycznymi. Jednocześnie wykazano, że tempo rozwoju narzędzi generatywnych przewyższa obecnie możliwości ich skutecznej detekcji.

W konkluzji przyjęto, że technologia deepfake stanowi istotne wyzwanie dla współczesnego społeczeństwa informacyjnego, wpływając na wiarygodność przekazu medialnego, zaufanie społeczne oraz bezpieczeństwo informacyjne. Konieczne jest zatem podejście interdyscyplinarne, łączące rozwój technologii wykrywających manipulacje, odpowiednie regulacje prawne oraz edukację medialną, aby ograniczyć negatywne konsekwencje tego zjawiska przy jednoczesnym zachowaniu jego potencjału użytkowego.

Słowa kluczowe

deepfake, dezinformacja, manipulacja, sztuczna inteligencja, wizerunek

This article was written as part of an international project entitled „Protection of Image in Copyright Law in the Digital Media Era - A Polish-Spanish Comparative Approach" within the framework of cooperation between the Faculty of Law and Administration at the War Studies University in Warsaw and the Faculty of Law at Pablo de Olavide University in Seville, Spain.

Wprowadzenie

Postępująca cyfryzacja życia społecznego oraz dynamiczny rozwój technologii sztucznej inteligencji doprowadziły do powstania nowych form przetwarzania i generowania informacji. Jednym z najbardziej złożonych i jednocześnie kontrowersyjnych przejawów tych zmian jest technologia deepfake. Zjawisko to w sposób istotny wpływa na współczesną przestrzeń informacyjną, podważając tradycyjne kryteria autentyczności przekazu wizualnego i dźwiękowego oraz redefiniując pojęcie wiarygodności treści cyfrowych.

W niniejszym artykule zastosowano metody jakościowe, obejmujące analizę i krytyczny przegląd literatury naukowej, raportów instytucjonalnych oraz opracowań dotyczących sztucznej inteligencji, dezinformacji i bezpieczeństwa informacyjnego. Metodę tę uzupełnia analiza porównawcza definicji oraz klasyfikacji deepfake'ów, co pozwala na wieloaspektowe ujęcie badanego zjawiska oraz identyfikację jego kluczowych cech i mechanizmów funkcjonowania.

Podjęta analiza koncentruje się wokół następujących pytań badawczych: jakie są podstawowe definicyjne i technologiczne cechy deepfake'ów?; jakie czynniki sprzyjają rozwojowi i rozpowszechnianiu technologii deepfake?; jakie zagrożenia wynikają z wykorzystania deepfake'ów w przestrzeni publicznej? oraz jakie są społeczne konsekwencje upowszechnienia treści syntetycznych?

Technologia deepfake wiąże się z licznymi zagrożeniami, wśród których szczególne znaczenie mają dezinformacja, naruszenia prywatności, manipulacja opinią publiczną oraz osłabienie zaufania do mediów i instytucji[1]. Jej wykorzystanie w kontekstach politycznych, ekonomicznych czy społecznych może prowadzić do destabilizacji procesów demokratycznych, eskalacji konfliktów oraz nowych form cyberprzestępczości. Jednocześnie deepfake stanowi wyzwanie dla systemów prawnych i etycznych, które nie zawsze nadążają za tempem rozwoju technologicznego.

Skutki społeczne upowszechnienia deepfake'ów wykraczają poza bezpośrednie szkody wynikające z pojedynczych nadużyć. Zjawisko to przyczynia się do erozji zaufania społecznego, relatywizacji prawdy oraz wzrostu niepewności informacyjnej. Odbiorcy coraz częściej kwestionują autentyczność materiałów audiowizualnych, co sprzyja chaosowi informacyjnemu i utrudnia racjonalną debatę publiczną. W tym kontekście analiza technologii deepfake staje się nie tylko zagadnieniem technologicznym, lecz także istotnym problemem społecznym, wymagającym interdyscyplinarnego podejścia[2].

Geneza technologii deepfake

Rozpatrując zagadnienia związane z deepfake, konieczne jest podkreślenie, że o ile sama technologia deepfake stanowi stosunkowo nowe

[1] Zob. szerzej: G. Pilarski, Cyberprzestrzeń – relacje w wojnie hybrydowej, Warszawa 2020.
[2] E. Bartuzi-Trokielewicz, Techniki manipulacji w świecie deepfake'ów – jak sztuczna inteligencja zmienia oblicze dezinformacji, https://ai.gov.pl/aktualnosci/Techniki-manipulacji-w-%C5%9Bwiecie-deepfake'%C3%B3w [dostęp: 1.12.2025].

zjawisko, o tyle podobne działania, obejmujące tworzenie fotomontaży oraz obróbkę materiału audiowizualnego, nie są niczym nowym i były często wykorzystywane w przygotowywaniu treści propagandowych oraz ukierunkowanych na manipulację percepcją odbiorcy. Postępujący rozwój technologiczny, związany z rozwojem Internetu oraz samouczącej się sztucznej inteligencji, stworzył natomiast nowe możliwości obróbki zdjęć, filmów oraz dźwięku, nie tylko skracając czas potrzebny do ich przygotowania, ale także zwiększając wiarygodność wygenerowanych produktów.

Po raz pierwszy termin deepfake pojawił się w 2017 r. i stanowił nawiązanie do użytkownika portalu Reddit, który opublikował fałszywe, ale niezwykle realistyczne filmy pornograficzne. Powstały one poprzez wkomponowanie w oryginalną treść filmu, twarzy znanych aktorek, m.in. Emmy Watson, Taylor Swift i Scarlet Johansson[3].

Wydarzenie to uznawane jest za początek ery deepfake, bowiem w niedługim czasie, w Internecie zaczęły pojawiać się nowe nagrania wygenerowane przy pomocy tej technologii, obejmujące zarówno etyczne (m.in. edukacyjne, rozrywkowe), jak i kontrowersyjne materiały (naruszające dobra osobiste, pornograficzne)[4]. Jednocześnie zaczęły powstawać nowe programy i aplikacje, umożliwiające użytkownikom – nawet nieposiadającym zaawansowanej wiedzy technologicznej – na generowanie materiałów na podstawie wgrywanych zdjęć, nagrań, czy próbek głosu, modyfikując je w taki sposób, aby były zgodne z oczekiwaniami ich twórców[5].

Z czasem, wraz z możliwościami jakie dawał postęp technologiczny, zakres wykorzystywania deepfake'ów uległ znacznemu zwiększeniu, podobnie jak i intensywność oraz częstotliwość ich wykorzystywania. Z technologii tej korzystali już nie tylko anonimowi użytkownicy Internetu, ale także zaczęła być ona wykorzystywana przez międzynarodowe korporacje, państwa, czy też inne organizacje, które w zależności od swoich intencji oraz zakładanych celów, generowały i rozpowszechniały treści w celu promocji swoich towarów lub idei[6], a także na potrzeby prowadzenia działań propagandowych, dezinformacyjnych lub dyskredytujących określone osoby lub podmioty[7]. W literaturze przedmiotu wskazuje się, że technologia deepfake została „zdemokratyzowana"[8].

Współcześnie, z uwagi na szeroki dostęp do licznych materiałów, zwłaszcza umieszczanych w otwartym dostępie, które mogą stanowić materiał źródłowy dla samouczących się algorytmów, a także rozwój technologiczny i stwarzane dzięki

[3] P. Galemba, Deepfake jako zagrożenie i wyzwanie dla bezpieczeństwa informacyjnego państwa, „Cybersecurity and Law" 2025, nr 1(13), s. 223.
[4] I. Dąbrowska, Deepfake – nowy wymiar internetowej manipulacji, „Zarządzanie Mediami" 2020, nr 2(8), s. 93-96.
[5] Zob. szerzej: M. Nowikowska, Deepfake w nowych mediach – zagrożenie dla tożsamości cyfrowej, „Roczniki Kulturoznawcze" 2025, vol. 17, nr 4, s. 272; K. Mularz, Deepfake jako generator fikcyjnych treści i rzeczywistej nieufności, „Dydaktyka Informatyki" 2024, nr 19, s. 61; G. Garcia Vera, Deep fake – postęp technologiczny a prawo karne, „Acta Iuridica Resoviensia" 2024, nr 1(144), s. 41.
[6] Zob. szerzej: Z. Spyra, M. Pyjas, Wykorzystanie narzędzi deepfake w komunikacji merketingowej a społeczna odpowiedzialność biznesu z perspektywy pokolenia Z – wyniki badań eksploracyjnych [w:] H. Mruka, A. Sawicki (red.), Marketing. Koncepcje i doświadczenia, Pelpin 2024, s. 148-149.
[7] Przykładem jest wykorzystanie przez Rosję nagrania deepfake w celu dyskredytacji prezydenta Ukrainy oraz skłonienie ukraińskich żołnierzy do dezercji i zaprzestania dalszej walki, P. Różyński, Deepfake idzie na wojnę, „Rzeczpospolita", 22 marca 2022, s. L2.
[8] H. Smith, K. Mansted, Weaponised Deep Fakes: National Security and Democracy, Australia: Australian Strategic Policy Institute 2020, s. 3.

niemu możliwości szczegółowej analizy i grupowania wgrywanych do programu danych oraz łączenia ze sobą różnych kategorii danych wejściowych (obraz, film oraz dźwięk), generowane produkty deepfake, cechują się nieprawdopodobną dokładnością oraz odwzorowaniem szczegółów, których odróżnienie od oryginału jest niemal niemożliwe[9]. Rozwój omawianej technologii prowadzi tym samym do powstania zjawiska nazywanego głęboko fałszywą rzeczywistością[10].

Pojęcie deepfake – etymologia i zakres znaczeniowy

Istota zjawiska deepfake, ukryta jest w jego nazwie, będąc połączeniem dwóch słów pochodzących z języka angielskiego - deep (głęboki – w znaczeniu głębokiego uczenia się), a także fake (fałszywy, spreparowany, podrobiony)[11]. W literaturze przedmiotu można odnaleźć wiele odmiennych ujęć i prób zdefiniowania tego terminu.

Pojęcie deepfake zostało uregulowane na gruncie Rozporządzenia Parlamentu Europejskiego i Rady 2024/1689 (tzw. aktu w sprawie sztucznej inteligencji)[12]. 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 Rozporządzenia 2024/1689 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. Zgodnie z dyspozycją art. 50 ust. 4 Rozporządzenia 2024/1689, 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[13].

Zjawisko deepfake stanowi także przedmiot analiz przedstawicieli polskiej i zagranicznej doktryny. 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[14]. Mette-Marie

[9] Z. Spyra, M. Pyjas, op.cit., s. 148.
[10] J. Bernacka, Problematyka prawna technologii deepfake – analiza legalności tworzenia i rozpowszechniania deepfake'ów po uchwaleniu AI Act, „Prawo i Więź" 2025, nr 5(58), s. 674.
[11] P. Galemba, op.cit., s. 223.
[12] Rozporządzenie 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 (Dz.Urz.UE L. z 12.7.2024).
[13] M. Łabuz, Deep fakes and the Artificial Intelligence Act - An important signal or a missed opportunity?, „Policy & Internet" 2024, nr 16, s. 11, https://doi.org/10/1002/poi3.406.
[14] N. Young, DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media, New York: Independently published 2019, s. 14.

Zacher Sørensen zauważa, że deepfake to fałszywy film (syntetyczna audiowizualna produkcja medialna) wyprodukowany przy użyciu technik głębokiego uczenia (sztucznych sieci neuronowych)[15]. Podobnie Jon Bateman podkreśla, że deepfake to media generowane przez sztuczną inteligencję, które przedstawiają wymyślone wydarzenia, czasami całkiem realistycznie[16].

Na gruncie doktryny polskiej Damian Flisak[17] wskazuje, że termin ten oznacza „nakierowane na wywołania negatywnych konsekwencji, cyfrowe wytworzenie, ingerencję lub manipulację rzeczywistym obrazem osób". Autor dalej zauważa, że powszechnie zajwisko to było powiązane z działaniem intencjonalnym. W rozpowszechnionym użytku cechą wskazanego terminu było celowe działanie autora deepfake nakierowane na wyrządzenie krzywdy innej osobie. Z tego punktu widzenia, zawarta w Rozporządzeniu 2024/1689 definicja jest znacznie szersza, gdyż jest uniezależniona od celu działania twórcy deepfake'u.

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, można stworzyć zupełnie nowe i całkowicie fikcyjne zachowanie danej osoby[18]. Oliwia Rajtar określa deepfake jako „technologię opartą na sztucznej inteligencji, pozwalającą na tworzenie realistycznych, ale fałszywych filmów poprzez łączenie i modyfikowanie istniejących obrazów i nagrań"[19]. Na uwagę zasługuje także definicja zawarta w Leksykonie Cyberbezpieczeństwa, w której określono, że „deepfake to rodzaj sztucznej inteligencji do tworzenia będących oszustwem realistycznych obrazów rzeczywistości"[20], jednocześnie wskazując, że „pozwala na stworzenie dowolnego materiału przedstawiającego dowolną osobę w sytuacji, która w rzeczywistości nigdy nie miała miejsca"[21].

Podsumowując powyższe można stwierdzić, że deepfake najczęściej odnosi się do sfabrykowanego filmu lub dźwięku przedstawiającego osobę mówiącą lub robiącą coś, czego nigdy nie powiedziała[22]. Jednocześnie, należy w tym miejscu podkreślić, że deepfake nie musi obejmować wyłącznie wizerunku osób, ale także może w sposób spreparowany przedstawiać określone fikcyjne wydarzenie lub miejsce, wykorzystywane w celu poparcia przekazywanych treści lub w celu zbudowania określonej reakcji u odbiorcy[23]. Analiza przedmiotowych definicji pozwala na wskazanie dwóch elementów charakteryzujących deepfake: procesu głębokiego uczenia się oraz fałszerstwa.

---

[15] M.M. Zacher Sørensen, Deepfake Face-Swap Animations and Affect, [w:] T. Tomoko (red.) Human Perception and Digital Information Technologies: Animation, the Body, and Affect, Bristol 2024, s. 195.

[16] J. Bateman, Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios, Waszyngton 2020, s. 4.

[17] D. Flisak, O niełatwym wytyczaniu granic rzeczywistości przez AI Act: zasada transparentności, deepfake, „Prawo Nowych Technologii" 2024, nr 1, s. 42.

[18] L. Wasilewski, Technologia deepfake wymaga regulacji – komputerowe przeróbki stają się zagrożeniem, https://www.prawo.pl/biznes/pierwsze-regulacje-prawne-technologii-deepfake.49454.html [dostęp 14.12.2024].

[19] O. Rajtar, Technologie deepfake w zakresie danych osobowych, „Rocznik Administracji Publicznej" 2025, nr 11(1), s. 194.

[20] K. Kaczmarek, Deepfake [w:] K. Chałubińska-Jentkiewicz (red.), Leksykon cyberbezpieczeństwa, Warszawa 2024, s. 97.

[21] Tamże, s. 97.

[22] J. Bateman, op.cit., s. 4.

[23] M. Nowikowska, Deepfake w nowych mediach..., s. 274.

Autorzy, na potrzeby niniejszego artykułu, proponują definicję omowianego pojęcia w ujęciu szerszym i węższym. Ujęcie sensu stricto (węższe) deepfake jest to „zastosowanie technik obróbki obrazu, dźwięku lub wideo z użyciem algorytmów sztucznej inteligencji, pozwalających realistycznie podmieniać twarze, głosy lub całe sceny tak, żeby wyglądały na autentyczne, mimo że zostały sztucznie wygenerowane, mogące prowadzić do błędnej oceny autentyczności treści w celu wprowadzenia odbiorcy w błąd".

Ujęcie sensu largo (szersze) „jest to zjawisko technologiczno-społeczne polegające na wytwarzaniu, dystrybucji i odbiorze syntetycznych treści audiowizualnych, które mogą prowadzić do zafałszowania rzeczywistości, w celu wywierania wpływu w obszarze społecznym, politycznym, prawnym, gospodarczym, informacyjnym i kognitywnym".

Podsumowując powyższe można stwierdzić, że tchnologię deepfake należy postrzegać z dwóch perspektyw: węższej, odnoszącej się technologii służącej do realistycznego fałszowania treści audiowizualnych, audio, video lub obrazu lub oraz szerszej, jako zjawisko technologiczno-społeczne polegające na wytworzeniu i wykorzystaniu treści syntetycznych w celu wywierania wpływu w obszarze społecznym, politycznym, prawnym, gospodarczym, informacyjnym i omawianego pojęcia.

Mechanizmy funkcjonowania deepfake

Generowanie materiałów deepfake oparte jest o technologię nazywaną generatywnymi sieciami przeciwstawnymi (ang. Generative Adversarial Networks – GAN), które składają się z dwóch wzajemnie przeciwstawnych algorytmów opartych o samouczącą się sztuczną inteligencję – generatora oraz dyskryminatora. Tworzenie materiałów deepfake polega na zaimportowaniu do specjalistycznego programu wyselekcjonowanych danych wejściowych, stanowiących bazę zawierającą próbki obrazów, dzięków oraz nagrań wideo, na bazie których ma zostać przygotowany ostateczny produkt. Na ich podstawie pierwszy z algorytmów wchodzących w skład GAN, tj. generator, analizuje wprowadzone dane oraz wykrywa charakterystyczne cechy wspólne, takie jak: ton głosu, mimika, gestykulacja, a następnie dokonuje ich segregacji i scalania, kreując tym samym oczekiwany materiał. Następnie, drugi z algorytmów – dyskrymnator – dokonuje weryfikacji otrzymanego produktu w celu określenia poziomu jego wiarygodności. Proces generowania oraz weryfikacji powtarza się do chwili, gdy dyskryminator uzna materiał za wiarygodny i zgodny ze wszystkimi cechami charakterystycznymi[24].

Wykorzystanie technologii GAN umożliwia nie tylko szybkie generowanie materiałów deepfake, ale także uzyskanie niezwykle wiarygodnych materiałów, które z uwagi na dbałość o szczegóły oraz opierające się o charakterystyczne cechy wspólne wszystkich wprowadzonych próbek (liczące często nawet kilka tysięcy) sprawiają, że ich weryfikacja oraz wykrycie bez zastosowania specjalistycznych narzędzi opartych o sztuczną technologię, jest niemal niemożliwa.

Podsumowując tę część rozważań można stwierdzić, że deepfake to technika syntezy obrazów oparta na sztucznej inteligencji. System ten składa się z dwóch sztucznych sieci: generatora i dyskryminatora. Sieci te pozostają w relacji konkurencyjnej. Generator tworzy fałszywe dane, dyskryminator

[24] Por.: P. Galemba, op.cit., s. 223, K. Mularz, op.cit., s. 57; Z. Spyra, M. Pyjas, op.cit., s. 148.

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[25]. Sieci konkurują ze sobą aż do momentu, w którym nie można już odróżnić prawdziwych wyników od fałszywych[26].

Charakterystyka zjawiska deepfake

Deepfake stanowi niezwykle złożone zjawisko, obejmujące szeroko pojęte możliwości jego przygotowywania oraz wykorzystywania. Z tego powodu, w celu właściwego zrozumienia jego istoty, przyjmuje się różnorodne typologie, oparte o kryterium wykorzystywanej technologii, jak i celu, w jakim mają zostać użyte wygenerowane produkty końcowe.

Pierwszą z kategoryzacji, jest przedstawiony przez Oskara Bodanka podział ze względu na zastosowaną technologię, według której podzielić deepfake można na dwa zasadnicze rodzaje: audio oraz wideo. Do pierwszego z nich Autor zalicza syntezę mowy (ang. speech synthesis), która polega na wygenerowaniu dźwięku opartego o próbki głosu określonej osoby (uwzględniając ton, akcent, intonację), a następnie odczytanie nim tekstu wprowadzonego do systemu przez użytkownika[27].

Drugim rodzajem deepfake, są materiały wideo, które ze względu na sposób tworzenia można podzielić na trzy kategorie:

face reenactment, czyli generowanie twarzy oraz nałożenie jej na twarz postaci (pozoranta) w oryginalnym nagraniu;

face swap, rozumiana jako wzajemna zamiana dwóch twarzy pomiędzy próbkami (zdjęciami, nagraniami) oraz

face generation, czyli generowanie całkowicie nowego, nieistniejącego w rzeczywistości, obrazu twarzy na podstawie wgranych próbek[28].

Iwona Dąbrowska dokonuje podziału technologii deepfake ze względu na cel, jaki przyświeca twórcy wygenerowanych treści. Autorka rozróżnia cztery kategorie:

rozrywkowe – obejmujące żartobliwe treści i wizerunki powszechnie znanych postaci z popkultury lub fikcyjne postacie;

edukacyjne – wykorzystywanie wizerunku lub głosu znanej postaci w celu przedstawienia określonych treści edukacyjnych;

dezinformacyjne – ukierunkowane na wzbudzeniu niepokoi społecznych lub wprowadzenia chaosu informacyjnego;

dyskredytacyjne – generowane w celu zdyskredytowania osoby, grupy społecznej lub innego podmiotu[29].

Oprócz powyższych kategorii, warto dodać jeszcze dwie, w których deepfake odgrywa coraz większą rolę, tj. marketing oraz pornografia. Przykładem pierwszego z nich bez wątpienia są spoty oraz całe kampanie reklamowe dużych koncernów i przedsiębiorstw, które wykorzystując omawianą technologię dążą do przygotowania jak najbardziej kreatywnych oraz przekonywujących reklam

[25] S. Mansted, op.cit., s. 8.
[26] M. Nowikowska, Deepfake w nowych mediach..., s. 274-275.
[27] O. Bodanka, Naruszenie wizerunku przy wykorzystaniu technologii deepfake – analiza prawna i praktyczna, „Opolskie Studia Administracyjno-Prawne" 2022, nr 1(20), s. 14.
[28] Ibidem, s. 15.
[29] I. Dąbrowska, Deepfake – nowy wymiar internetowej manipulacji, „Zarządzanie Mediami" 2020, t. 8(2), s. 96.

swoich produktów. Przykładem może być Pepsi, która w swojej reklamie wykorzystała motyw spotkania aktora z wygenerowaną młodszą wersją siebie, czy też Nike, które wykorzystało deepfake, w celu ukazania rozgrywki pomiędzy współczesnymi oraz legendarnymi postaciami ze świata piłki nożnej[30].

Inną kategorią wykorzystywania deepfake, którą można uznać za nieetyczną, jest wykorzystywanie technologii deepfake na potrzeby generowania filmów pornograficznych, opierających się nie tylko na podłożeniu twarzy określonej osoby (jak miało to miejsce w przypadku pierwszego użycia deepfake), ale również w zakresie generowania całych postaci. Materiały tego rodzaju, określane terminem deepporn, zyskują coraz większą popularność w mediach społecznościowych, jak również na platformach umożliwiających dostęp wyłącznie subskrybentom, generując tym samym twórcy oraz osobom rozpowszechniającym te materiały dochody finansowe[31].

Jednocześnie warto w tym miejscu podkreślić, że tendencje wykorzystywania technologii deepfake będą wstale wzrastać, co wynika przede wszystkim ze stosunkowo krótkiego czasu ich przygotowywania oraz uzyskiwania wysokich efektów w zakresie wiarygodności generowanych materiałów. Jednocześnie aplikacje i programy wykorzystywane do ich realizacji cechują się dosyć dużą intuicyjnością oraz łatwością w obsłudze, nawet przez osoby niemające doświadczenia oraz specjalistycznej wiedzy w zakresie sieci generatywnych[32]. Z uwagi na wysoką atrakcyjność generowanych treści, coraz częściej użytkownicy Internetu oraz inne podmioty sięgają po tego rodzaju rozwiązania[33], prowadząc tym samym do zwiększenia możliwości samouczenia się algorytmów, które na bazie ogólnodostępnych fotografii, wideo oraz zapisów dźwiękowych[34], a także próbek wprowadzanych do bazy danych, zyskują nowe możliwości w zakresie ich grupowania i analizowania, a także dopierania i weryfikowania cech wspólnych.

Deepfake a inne formy manipulacji cyfrowej

Dynamiczny rozwój technologii informatycznych, szczególnie sztucznej inteligencji i uczenia maszynowego, uaktywnił nowy rodzaj zagrożeń informacyjnych – manipulacje cyfrowe. Współczesny odbiorca funkcjonuje w świecie, w którym granica między rzeczywistością a cyfrową fikcją zaczyna się zacierać. Zjawisko deepfake stanowi jedynie najbardziej zaawansowaną formę znacznie szerszej palety technik manipulacji obrazem, dźwiękiem i tekstem.

Równolegle z rozwojem zaawansowanych technik opartych na AI, wciąż rozpowszechnione są także prostsze formy manipulacji, określane często mianem cheapfakes lub shallowfakes[35] (tzw. tanie fałszerstwa). Są to tradycyjne (ręczne) metody obróbki graficznej, które nie wymagają stosowania sztucznej inteligencji. Jako przykład można wskazać fotomontaż, czy edycję selektywną, polegającą na zmianie kolorów, wycinaniu fragmentów zdjęć, co może prowadzić do zmiany kontekstu fotografii. Cheapfakesy, choć technicznie wydaje się mniej

[30] Zobacz: Z. Spyra, M. Pyjas, op.cit., s. 148-149.
[31] K. Mularz, op.cit., s. 59-60.
[32] M. Nowikowska, M. Kozłowski, Analiza wybranych zagrożeń dla tożsamości cyfrowej wynikających z postępu technologicznego, „Journal of Modern Science", 2025, nr 3(63), s. 792.
[33] J. Bernacka, op.cit., s. 675.
[34] M. Nowikowska, M. Kozłowski, op.cit., s. 793.
[35] J. R. Ostrowski, Shallowfakes, „The New Atlantis" 2023, no. 72, s. 96; https://www.jstor.org/stable/27212358. [dostęp: 1.12.2025].

wyrafinowane, mogą być równie szkodliwe. Wystarczy odpowiednio zmanipulowane zdjęcie lub krótkie nagranie, by wywołać społeczne oburzenie, wzbudzić fałszywe przekonania lub zaszkodzić czyjemuś wizerunkowi. Płytkie fałszerstwa charakteryzują się niższą jakością edycji obrazu i wideo w porównaniu z wysokiej klasy tworzeniem głębokich fałszerstw. Stosunkowo płytkie fałszerstwa można łatwo tworzyć za pomocą podstawowych narzędzi internetowych, a także łatwo je rozpoznać[36].

Istotnym elementem współczesnej manipulacji cyfrowej są również generowane przez sztuczną inteligencję obrazy, które nie stanowią przeróbki istniejących materiałów, lecz powstają całkowicie od podstaw. Modele takie jak Midjourney czy Stable Diffusion mogą generować realistyczne obrazy od zera. Za pomocą specjalistycznych modeli można stworzyć realistyczne fotografie przedstawiające wydarzenia, które nigdy nie miały miejsca, lub osoby, które nie istnieją. W erze mediów społecznościowych tego typu wytwory mogą rozprzestrzeniać się błyskawicznie, utrudniając odbiorcom ocenę autentyczności treści[37].

Kolejną formą manipulacji cyfrowej jest zjawisko powszechnie zwane klonowaniem głosu[38]. Coraz powszechniejsze stają się techniki klonowania głosu, umożliwiające tworzenie syntetycznych nagrań audio imitujących sposób mówienia konkretnych osób. Agata Ziobroń wskazuje, że klonowanie głosu (audio deepfake) „polega na przekonującym przetworzeniu głosu, w efekcie którego nadaje się mu cechy głosu innej osoby”[39]. Technika ta została opracowana pierwotnie w pozytywnych celach, takich jak tworzenie audiobooków poprzez imitację kojącego głosu. Podobną rolę odgrywają generatory tekstu, które potrafią tworzyć pozornie wiarygodne artykuły prasowe, komentarze, recenzje czy wypowiedzi w mediach społecznościowych. Automatyczne generowanie treści ułatwia produkcję fake newsów, pozwala tworzyć masowe kampanie propagandowe i zwiększa trudność w identyfikacji autentycznych źródeł informacji.

Na szczególną uwagę zasługują także manipulacje kontekstowe[40], które nie wymagają ingerencji technicznej w materiał, lecz opierają się na zmianie sposobu jego przedstawienia. Przykładem może być wycięcie fragmentów wypowiedzi, tak aby zmienić ich sens, lub opatrzenie prawdziwego zdjęcia fałszywym opisem. Tego rodzaju zabiegi są często trudniejsze do wykrycia, ponieważ w oczywisty sposób nie naruszają struktury wizualnej materiału.

Wraz z rozwojem technologii powstają również zautomatyzowane systemy manipulacji – sieci botów oraz syntetyczne postacie, które działają w mediach społecznościowych. Boty te, wyposażone w zdjęcia generowane przez AI i w pełni sztuczne profile, są w stanie prowadzić spójne rozmowy, udostępniać

[36] A. Gibson, Digital Humanities in the Deepfake Era, [in:] M.K. Gold, L.F. Klein (eds.), Debates in the Digital Humanities, University of Minnesota Press 2023, s. 166, http://www.jstor.org/stable/10.5749/j.ctv345pd4p.13 [dostęp: 1.12.2025].

[37] K. Giles, K. Hartmann, M. Mustaffa, The Role of deepfakes in mailing influence campaigns, Riga: NATO Strategic Communications Centre of Excellence 2019, s. 8.

[38] Zob. szerzej: A. Ziobroń, Klonowanie głosu - wyzwaniem dla prawa karnego?: O przestępczości z użyciem sztucznej inteligencji, [w:] J. Piskorski, M. Błaszczak (red.), Nowe technologie: wyzwania i perspektywy dla prawa karnego, Łódź 2023, s. 241-256.

[39] Ibidem, s. 242.

[40] Por. K. Cymanow-Sosin, Perswazja i manipulacja jako zjawiska kontekstowe na przykładzie reklam sektora usług instytucji rynku finansowego, „Naukowy Przegląd Dziennikarski” 2015, nr 2, s. 52; J. Szczepańska-Włoch, (Nie)racjonalność jako strategia manipulacji. Wywiad polityczny w Wielkiej Brytanii, „Res Rhetorica” 2023, nr 10(2), s. 76-77.

treści i wpływać na opinie użytkowników[41]. Zacierają one granicę między rzeczywistym użytkownikiem a cyfrową imitacją, tworząc iluzję społecznego poparcia lub sprzeciwu[42].

Podsumowując powyższe można uzanć, że skala i złożoność nowoczesnych manipulacji cyfrowych prowadzą do fundamentalnych pytań o wiarygodność informacji oraz bezpieczeństwo społeczne. W świecie, w którym technologia może dowolnie kształtować obraz rzeczywistości, kluczowe staje się budowanie nowych kompetencji medialnych i rozwijanie narzędzi umożliwiających identyfikację fałszerstw. Zrozumienie natury deepfake'ów i pokrewnych technik jest pierwszym krokiem do świadomego funkcjonowania w przestrzeni cyfrowej, gdzie nie wszystko, co widzimy i słyszymy, musi być prawdziwe.

Potencjalne obszary nadużyć przez technologię deepfake

Technologia deepfake, oparta na zaawansowanych modelach sztucznej inteligencji, otworzyła przed społeczeństwem zupełnie nowe możliwości tworzenia realistycznych treści audiowizualnych. Choć w wielu dziedzinach jej rozwój ma charakter pozytywny, od przemysłu filmowego po edukację, to jednak równolegle ujawniły się liczne obszary nadużyć. Wykorzystanie deepfake'ów w celach szkodliwych, manipulacyjnych lub przestępczych stało się jednym z najpoważniejszych wyzwań współczesnego ekosystemu informacyjnego. Osoby wykorzystujące technologię deepfake mogą używać rozmaitych danych dotyczących człowieka, jego wizerunku, danych osobowych, ingerować w sferę dóbr osobistych, a także działać w różnych celach, często nieetycznych, prowadzących do pokrzywdzenia kogoś. To właśnie dlatego technologia deepfake może budzić kontrowersje.

Jednym z najbardziej niebezpiecznych obszarów nadużyć jest przestrzeń polityki i życia publicznego. Deepfake'i mogą być wykorzystywane do tworzenia materiałów przedstawiających polityków lub liderów opinii w sytuacjach, które nigdy nie miały miejsca, wygłaszających kontrowersyjne opinie, popełniających błędy lub zachowujących się w sposób skandaliczny. W okresach napięć społecznych czy kampanii wyborczych, takie nagrania mogą wywoływać chaos informacyjny, wpływać na decyzje wyborców lub podważać zaufanie do instytucji[43]. Z uwagi na ich potencjalnie destabilizujący charakter, deepfake'i są często postrzegane jako narzędzie nowoczesnych operacji informacyjnych, wykorzystywanych zarówno przez osoby fizyczne, organizacje, jak i państwa[44].

W praktyce jednym z najczęstszych nadużyć technologii deepfake jest generowanie materiałów pornograficznych z udziałem osób, które nigdy nie brały udziału w takich produkcjach. Zjawisko to dotyczy zarówno osób publicznych, jak i prywatnych, co czyni je wyjątkowo dotkliwym naruszeniem prywatności

[41] R. Ćwiertniak, Ludzie i boty sztucznej inteligencji w nowoczesnych projektach, „Studia i Prace. Kolegium zarządzania i finansów" 2024, nr 199, s. 83, https://doi.org/10.33119/SIP.2024.199.4.

[42] Por. A. Łukasik-Turecka, Boty w służbie dezinformacji, „Polityka i Społeczeństwo" 2024, nr 4(22), s. 137, DOI: 10.15584/polispol.2024.4.9; M. Himelein-Wachowiak, S. Giorgi, A. Devoto, M. Rahman, L. Ungar, H.A. Schwartz, D.H. Epstein, L. Leggio, B. Curtis, Bots and Misinformation Spread on Social Media: Implications for COVID-19, „Journal of Medical Internet Research" 2021, vol. 23, no. 5, doi: 10.2196/26933;

[43] K. Giles, K. Hartmann, M. Mustaffa, op.cit., s. 8.

[44] M. Nowikowska, PSYOPS as an elemnt of information warlfare, [in:] K. Chałubińska-Jentkiewicz, O. Evsyukova, Information, disinformation, cybersecurity, Toruń 2023, s. 163 i n.

i autonomii jednostki[45]. Materiały takie mogą być wykorzystywane do szantażu, nękania lub poniżenia, a ich rozpowszechnianie bywa niezwykle trudne do zatrzymania. W skrajnych przypadkach deepfake'i przeradzają się w cyfrową przemoc, której skutki psychologiczne są równie realne jak w przypadku tradycyjnych form nękania.

Deepfake'i audio, oparte na syntetycznym klonowaniu głosu, znalazły zastosowanie w nowych formach wyłudzeń. Cyberprzestępcy wykorzystują wygenerowane nagrania, aby podszywać się pod przełożonych, współpracowników, czy członków rodziny, nakłaniając ofiary do przelewów czy ujawnienia poufnych danych. Ten rodzaj oszustw stał się szczególnie niebezpieczny w środowisku biznesowym, gdzie głos przełożonego czy prezesa bywa traktowany jako wiarygodne źródło poleceń. Deepfake może więc funkcjonować jako narzędzie inżynierii społecznej, łącząc perswazję psychologiczną z zaawansowaną technologią imitacji[46].

Paradoksalnie, deepfake'i mogą wyrządzać szkody nawet wtedy, gdy nie zostały użyte. Sam fakt istnienia technologii prowadzi do tzw. efektu odrzucenia rzeczywistości (ang. liar's dividend), polegającego na tym, że osoby winne nieetycznych działań mogą twierdzić, że autentyczne nagrania są fałszywe. Zjawisko to podważa wiarygodność materiałów wizualnych jako dowodów w życiu publicznym, mediach, a nawet w postępowaniu sądowym. Deepfake'i zmieniają więc nie tylko treść komunikacji, ale również sposób, w jaki człowiek postrzega prawdę i weryfikuje informacje. Grzegorz Pilarski trafnie zauważa, że obecnie deepfake, jako podstawowe narzędzie dezinformacji, wpływa na zachwianie fundamentalnego założenia wiarygodności przekazów audiowizualnych, co doprowadza do sytuacji, w której nawet autentyczne nagrania mogą być kwestionowane jako potencjalnie zmanipulowane[47]. Ma to odzwierciedlenie w zakresie zaufania społecznego i informacyjnego wpływając między innymi na normalizację nieufności informacyjnej, emocjonalną polaryzację społeczną, erozję reputacji jednostek zaufania publicznego, wzrost zagrożeń psychologicznych. Podobnie Ewelina Bartuzi-Trokielewicz wskazuje, że technologia deepfake to nowoczesna i kompleksowa broń psychologiczna[48]. Spójnie łączy iluzję wizualną z precyzyjną manipulacją emocjami i zaufaniem. Dlatego tak ważna jest edukacja społeczeństwa. W środowisku cyfrowym, gdzie treści rozprzestrzeniają się błyskawicznie, deepfake'i mogą być wykorzystywane do kreowania fałszywych kontrowersji, oczerniania osób publicznych lub wywoływania masowych emocji. Algorytmy platform społecznościowych premiują materiały wzbudzające silne reakcje, co sprzyja ekspansji zmanipulowanych nagrań. Fałszywe treści mogą stać się elementem kampanii dezinformacyjnych, działań marketingowych lub zwykłego internetowego nękania. Dynamiczna

[45] Zob. szerzej: A. Ziobroń, Deepfake a prawo karne. Uwagi de lege lata i de lege ferenda dotyczące fałszywej pornografii, „Studenckie Prace Prawnicze, Administratywistyczne i Ekonomiczne" 2021, nr 37, s. 226; M. Sewastianowicz, Deepfake - ofiara realistycznej przeróbki może mieć problem z dochodzeniem swoich praw, https://www.prawo.pl/prawo/deepfake-a-prawo-karne-film-porno, 520138.html, [dostęp: 1.12.2025]; J. Nabiałek, Wielka Brytania: Porno deepfake to przestępstwo od momentu utworzenia, https://forsal.pl/gospodarka/cyfryzacja/artykuly/9489216,wielka-brytania-porno-deepfake-to-przestepstwo-od-momentu-utworzenia.html [dostęp: 1.12.2025].

[46] A. Sanocki, Deepfakes, czyli postprawda objawiona, [w:] M. Wrzosek (red.), Zjawisko dezinformacji w dobie rewolucji cyfrowej. Państwo. Społeczeństwo. Polityka. Biznes, Warszawa 2019, s. 21; M. Nowikowska, Deepfake w nowych mediach..., s. 274.

[47] G. Pilarski, Cyberprzestrzeń – relacje w wojnie hybrydowej, Warszawa 2020.

[48] E. Bartuzi-Trokielewicz, op.cit.

natura mediów społecznościowych utrudnia zarówno wykrywanie deepfake'ów, jak i ograniczanie ich wpływu.

Kolejnym obszarem nadużyć przez technologię deepfake jest naruszenie prywatności i reputacji osób fizycznych. Technologia ta pozwala na tworzenie scen, w których podmiot zostaje umieszczony w sztucznie skonstruowanym kontekście, np. w materiałach kompromitujących lub obraźliwych. Powoduje to, że deepfake przyczynia się do redefinicji pojęcia „wizerunku". W przypadku podmienienia twarzy aktorki z filmu pornograficznego na twarz innej osoby mamy do czynienia z wykorzystaniem wizerunku, ponieważ twarz stanowi cechę człowieka, która jest dostrzegalna, fizyczna i tworzy jego wygląd[49].

Wreszcie deepfake'i stanowią potencjalne zagrożenie dla bezpieczeństwa narodowego. Mogą być wykorzystywane do symulowania wypowiedzi przywódców państw, fałszywych komunikatów wojskowych, alarmów kryzysowych czy działań przestępczych. W skrajnych scenariuszach mogłyby doprowadzić do paniki społecznej, eskalacji konfliktów lub nieporozumień między państwami. Wymaga to od instytucji państwowych tworzenia nowych protokołów weryfikacji autentyczności materiałów i oficjalnych komunikatów.

Podsumowując powyższe należy stwierdzić, że technologia deepfake, choć fascynująca i pełna potencjału, niesie ze sobą poważne ryzyka wielu nadużyć. Jej wpływ na społeczeństwo wykracza daleko poza przestrzeń technologiczną, dotyka fundamentów zaufania, wiarygodności oraz bezpieczeństwa informacji[50]. Zrozumienie obszarów, w których deepfake'y mogą być wykorzystywane w sposób szkodliwy, jest kluczowym krokiem w budowaniu odporności społeczeństwa na współczesne zagrożenia cyfrowe.

Znaczenie deepfake w kontekście zagrożeń społecznych

Działalność różnych użytkowników cyberprzestrzeni wykorzystujących zjawisko deepfaku może mieć ogromny wpływ w kontekście społecznym. Technologia deepfake od początku swojego oddziaływania[51] wywiera istotny wpływ na funkcjonowanie współczesnych społeczeństw w kontekście wiarygodności postrzegania informacji audiowizualnych. Klasyczny przekaz informacji[52] w postaci obrazu i dźwięku był uznawany za jeden z najbardziej zaufanych nośników informacji. Rozpowszechnienie technologii deepfaku zaburzyło ten proces, doprowadzając do kryzysu zaufania społecznego wobec mediów i komunikacji masowego przekazu[53].

Z punktu widzenia społecznego należy mieć świadomość, że deepfake jest zagrożeniem wynikającym z zamierzonych i konsekwentnych formuł przekazu informacji w postaci dezinformacji. Jest to obecnie temat największego zainteresowania, zarówno strony siejącej dezinformację, jak i strony broniącej się

[49] Zob. E. Wojnicka, Prawo do wizerunku w ustawodawstwie polskim, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego. Prace z Wynalazczości i Ochrony Własności Intelektualnej" 1990, z. 56, s. 107; J. Sieńczyło-Chlabicz, Przedmiot, podmiot i charakter prawa do wizerunku, „Przegląd Ustawodawstwa Gospodarczego" 2003, nr 8, s. 20

[50] Zob. K. Chałubińska-Jentkiewicz, M. Nowikowska, Ochrona informacji w cyberprzestrzeni, Warszawa 2020, s. 19.

[51] P. Galemba, op.cit., s. 223.

[52] G. Pilarski, Wirtualne relacje – szanse i zagrożenia wynikające z „życia" w cyberprzestrzeni, [w:] I. Horzela, A. Nowakowska-Krystman (red.), Zarządzanie relacjami wybrane aspekty, Warszawa 2015, s. 137-138.

[53] G. Pilarski, Cyberspace as a Tool of Contemporary Propaganda, „Safety & Defense" 2020, tom 6 nr. 1, s. 101, https://doi.org/10.37105/sd.72, s. 101.

przed nią. W tym obszarze należy zwrócić szczególną uwagę na trzy zasadnicze kwestie. Po pierwsze, przekaz informacji. W tym zakresie należy rozpatrywać w jaki sposób i jakimi narzędziami dochodzi do przekazu informacji, czyli wykorzystaniu przede wszystkim cyberprzestrzeni, Internetu oraz mediów społecznościowych. Po drugie, należy rozważyć, czy działania zamierzone w postaci konsekwentnie przekazywanej dezinformacji są ze sobą w jakiś sposób powiązane? W ostatniej, trzeciej kwestii, należy zweryfikować, jaki wpływ ma dezinformacja na funkcjonowanie społeczeństwa.

W pierwszym obszarze należy podkreślić, że cyberprzestrzeń[54] - jako środowisko informacyjne - umożliwia przekazywanie informacji. Należy podkreślić, że cyberprzestrzeń jest złożonym i dynamicznie zmieniającym się środowiskiem, które wykorzystywane jest przez wszystkich, czyli „dobrą" i „złą" grupę użytkowników. Kwestie te zostały poruszone szerzej przez Grzegorza Pilarskiego w monografii pt. „Cyberprzestrzeń – relacje w wojnie hybrydowej". Autor zaproponował podział na grupę realizującą cyberkooperencję pozytywną oraz cyberkooperencję negatywną[55]. Pojęcie cyberkooperencji zdefiniował jako prowadzenie lub przygotowanie aktywności w cyberprzestrzeni, przy udziale podmiotów (aktorów) pozostających w jednoczesnych i współzależnych relacjach konkurencji i kooperacji w określonym horyzoncie czasowym[56].

W przedmiotowym przypadku, czyli szerzeniu dezinformacji lub też obronie przed nią, przyjmując za kryterium skutek oddziaływania, możemy wyodrębnić jednoznacznie działania negatywne oraz pozytywne. Autor dalej stawia pytanie, jaka jest różnica pomiędzy obiema grupami? Jeżeli przyjmiemy, iż funkcjonowanie podmiotów w cyberprzestrzeni będzie się odbywało bez znamion cyberprzestępstwa, tj. iż będzie to etyczne działanie w cyberprzestrzeni, niepowodujące żadnego negatywnego wpływu na funkcjonowanie innych podmiotów, z poszanowaniem życia prywatnego i komunikowania się, to przy takim założeniu podejmowane w cyberprzestrzeni aktywności będą się kwalifikowały do typu pozytywnego. Natomiast działania zgoła odmienne, nieetyczne i prowadzące do negatywnych efektów, będą zakwalifikowywane do aktywności negatywnych.

W rozpatrywanym przypadku, należy zwrócić szczególną uwagę na wykorzystanie portali społecznościowych, które mogą być używane zarówno przez agresora, jak i defensora. Należy podkreślić, że media społecznościowe odgrywają kluczową rolę, gdyż za ich pośrednictwem (jako podstawowe kanały komunikacji) mogą być osiągane przykładowo zamierzone cele polityczne, wojskowe, chociażby poprzez wpływanie na poglądy i zachowanie ludzi. Media społecznościowe, to z jednej strony narzędzie techniczne służące do komunikacji, np. X, ale także narzędzie do wymiany informacji, na przykład przekazanie nieprawdziwej wiadomości o śmierci głowy państwa. To z kolei może mieć ogromne znaczenie na prawidłowe funkcjonowanie państwa na płaszczyźnie politycznej, społecznej i ekonomicznej.

---

[54] W Polsce definicja cyberprzestrzeni została zamieszczona w nowelizowanej ustawie o stanie wojennym oraz kompetencjach Naczelnego Dowódcy Sił Zbrojnych i zasadach jego podległości konstytucyjnym organom RP z dnia 30 sierpnia 2011 r. Dokument ten określa, że cyberprzestrzeń to „przestrzeń przetwarzania i wymiany informacji tworzona przez systemy teleinformatyczne, określone w art. 3 pkt. 3 ustawy z dnia 17 lute-go 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne (Dz. U. z 2014 r. poz. 1114), wraz z powiązaniami pomiędzy nimi oraz relacjami z użytkownikami"

[55] G. Pilarski, Cyberprzestrzeń – relacje w wojnie hybrydowej, Warszawa 2020.

[56] Ibidem, s. 45.

Z uwagi na możliwości odtwarzania informacji w szybkim tempie przy małym koszcie, jak również trudności uznania, które informacje są rzetelne, a które sfabrykowane, media społecznościowe mogą być skutecznie stosowane do osiągania celów wojskowych. Do podstawowych zadań, jakie spełniają media społecznościowe, zalicza się: kształtowanie opinii publicznej, możliwość mobilizacji ludzi popierających jakąś ideę, możliwość koordynacji i synchronizacji działań wojskowych, wreszcie możliwość zbierania informacji, dzięki którym można ustalać cele ataku. Interesujące ujęcie wykorzystania mediów społecznościowych przedstawił Thomas E. Nissen w opracowaniu pt. „The Weaponization of Social Media”[57]. Autor wykazał, że wykorzystanie mediów społecznościowych może być traktowane jako broń w walce. Przy użyciu mediów społecznościowych możemy wyróżnić grupę działań, takich jak: targeting, zbieranie informacji, obrona, dowodzenie, operacje, jak również możemy dokonywać informowania i wpływania na społeczeństwo (propaganda), w zakresie kształtowania opinii, manipulowania, wywierania wpływu, zwodzenia czy przekonywania.

Przechodząc do drugiego obszaru, dotyczącego celu działania, w postaci konsekwentnie przekazywanej dezinformacji, należy zadać pytanie o wzajemne powiązania, należy rozważyć, czy działania zamierzone w postaci konsekwentnie przekazywanej dezinformacji są ze sobą w jakiś sposób powiązane? Dezinformacja jest elementem propagandy, gdzie narracja jest fałszywa. Największym zagrożeniem jest dezinformacja zarządzana, ukierunkowana na osiągnięcie zamierzonych celów długoterminowych. W tym przypadku propaganda będzie celem, a dezinformacja metodą jego realizacji. W zakresie analizy przypadków propagandy i perswazji w przestrzeni publicznej, Garth S Jowett oraz Victorie O'donnell w publikacji pn. „Propaganda & Persuasion”[58] przedstawili model propagandy jako formy komunikacji. Autorzy opisują m.in. proces przekazywania informacji i perswazję, cechy propagandy i jej mechanizmy jako procesu komunikacji z wykorzystaniem mediów społecznościowych. Przykład analizy, zgodnie z tym modelem, przedstawił Grzegorz Pilarski w artykule pn. „Cyberspace as a Tool of Contemporary Propaganda”[59], w którym dokonano identyfikacji działań propagandowych opartych na zaangażowaniu Republiki Niemiec w zakresie kazusu uchodźców z Syrii. Ten przykład doskonale wskazuje na znaczenie i możliwości, jakie daje cyberprzestrzeń w realizacji przedmiotowych działań i osiąganiu wyznaczonych celów. Należy podkreślić, że cyberprzestrzeń - ze wszystkimi jej elementami - jest warunkiem sukcesu działań propagandowych i dezinformacyjnych prowadzonych w dzisiejszych czasach.

Nieodłącznie z dezinformacją oraz zastosowaniem mediów społecznościowych wiąże się tzw. zjawisko „internet trolling”, którego celem jest kreowanie opinii publicznej poprzez zastosowanie fałszywych tożsamości w mediach społecznościowych. W działaniach tych wykorzystuje się takie metody, jak manipulowanie ludźmi dzięki rozpowszechnianiu propagandy, nieprawdziwych informacji, włączanie się do dyskusji z agresywnymi komentarzami. Zgodnie z podziałem przyjętym przez NATO Strategic

[57] Thomas T.E. Nissen, The Weaponization of Social Media, Royal Danish Defence College, Copenhagen 2015.
[58] G.S. Jowett, V. O'Donnell, Propaganda & Persuasion. London 2018.
[59] G. Pilarski, Cyberspace as a…, s. 95.

Communications Center of Excellence[60] wyróżnia się dwa rodzaje trollingu: klasyczny - który działa „dla zabawy", prowokując, siejąc konflikty w Internecie, bez jasno określonej agendy politycznej lub kierowania przez jakiś aktorów państwowych oraz hybrydowy - który działa jako część operacji hybrydowych, realizując cel polityczny lub państwowy. Hybrid troll jest wykorzystywany jako narzędzie prowadzenia wojen informacyjnych (informacyjnych kampanii), często pod kierowaniem lub na zlecenie konkretnego państwa lub organizacji. Trolling hybrydowy jest poważnym zagrożeniem, który może być wykorzystany w działaniach dezinformacyjnych, co niewątpliwie ma duże znaczenie z punktu widzenia społecznego.

W kolejnym obszarze należy zwrócic uwagę na działania dezinformacyjne, a w szczególności deepfake'i, które są bardzo niebezpieczne. Działanie te, jeżeli są długofalowe, mogą mieć kluczowy wpływ na bezpieczeństwo państwa, odporność społeczną oraz funkcjonowanie instytucji demokratycznych. Skuteczność dezinformacji wynika z połączenia trzech elementów:

1) skalowalności technologicznej, czyli zdolności współczesnych narzędzi cyfrowych do masowego, automatycznego i natychmiastowego rozpowszechniania treści dezinformacyjnych, poprzez zastosowanie automatyzacji, sieci botnetów i algorytmów profilowania behawioralnego;
2) precyzyjnego targetowania emocjonalnego, poprzez wykorzystywanie wiedzy o emocjach, lękach, uprzedzeniach i podatnościach odbiorców do dopasowywania treści w sposób maksymalizujący ich wpływ;
3) zdolności do trwałego zaburzania procesów poznawczych odbiorców, czyli najbardziej długofalowy i niebezpieczny element. Dezinformacja może wpływać na sposób, w jaki ludzie interpretują świat, podejmują decyzje i budują opinie.

Podsumowując powyższe należy stwierdzić, że dezinformacja może mieć poważne skutki w wielu obszarach naszego funkcjonowania, w tym społeczne, polityczne, dla bezpieczeństwa, gospodarcze oraz psychologiczne.

Skutki społeczne przejawiają się m.in. w polaryzacji i podziale społeczeństwa, erozji zaufania społecznego, jak również w konsekwencji normalizacji fałszu. Skutki polityczne widoczne są poprzez chociażby destabilizację procesów demokratycznych, wzmacnianie ekstremizmów, czy też umożliwianie zewnętrznych operacji wpływu. Skutki dla bezpieczeństwa przejawiają się poprzez wpływ na osłabienie odporności poznawczej społeczeństwa, wpływanie na destabilizację w sytuacjach kryzysowych, potęgując odczucia społeczne, a także ułatwienie działań wrogich aktorów. Skutki gospodarcze mogą objawiać się poprzez straty finansowe, zakłócenie rynków, czy też mogą wpływać na rosnące koszty przeciwdziałania dezinformacji. Natomiast skutki psychologiczne mogą obejmować zaburzenia procesów poznawczych, wzrost lęku, gniewu i bezradności, a także mogą powodować bardzo niebezpieczne zjawisko tzw. zmęczenia informacyjnego lub przesytu sprzecznych komunikatów, co może powodować, że odbiorcy przestają analizować treści i wybierają informacje zgodne z emocjami, nie z faktami.

W okresie długoterminowym, dezinformacja może skutkować chociażby powstawanie alternatywnych „rzeczywistości informacyjnych", osłabieniem więzi społecznych, a także powodować trwałe szkody w systemie politycznym i gospodarczym.

[60] NATO Strategic Communications Center of Excellence, Social Media as a Tool of Hybrid Warfare, Riga 2016, https://www.stratcomcoe.org/social-media-tool-hybrid-warfare (dostęp: 10.12.2025).

Konkludując, celowe działania dezinformacyjne mogą prowadzić do chaosu informacyjnego, destabilizacji instytucji, erozji zaufania społecznego, strat gospodarczych i trwałego osłabienia bezpieczeństwa państwa i społeczeństwa. Innymi słowy dezinformacja może podważyć fundamenty społeczne, polityczne i ekonomiczne, na których opiera się stabilne państwo. Istotnym walorem dezinformacji jest fakt, że aktorzy cyberkooperencji negatywnej wykorzystują działania dezinformacyjne, osiągając swoje cele bez konieczności użycia sił zbrojnych, w klasycznym rozumieniu, poprzez wojnę informacyjną, jako skuteczną broń o niskich kosztach i nieograniczonym zasięgu. Technologia deepfake może być bardzo skutecznym i niebezpiecznym narzędziem dezinformacji.

Wyzwania związane z identyfikacją i detekcją deepfake

Deepfake'i przenikły współcześnie do wszystkich aspektów życia społecznego, a ich oddziaływanie zależne jest od woli nie tylko osoby generującej dane treści, ale także osób, które później będą je rozpowszechniać w przestrzeni informacyjnej. Z tego powodu, identyfikacja oraz właściwa weryfikacja treści opartych na deepfake'ach, stanowi szczególne wyzwanie dla bezpieczeństwa państwa. Należy podkreślić, że nie jest to zadanie proste, bowiem z uwagi na nierozerwalny związek deepfake'ów z cyberprzestrzenią, stanowiącą stosunkowo trudne środowisko w zakresie wykrywania i zwalczanie działań propagandowych[61], to wnikliwa analiza treści, bez wykorzystania specjalistycznych narzędzi, staje się niemal niemożliwa dla przypadkowego odbiorcy. Dodatkowym elementem utrudniającym ich weryfikację jest wysoka wiarygodność oraz dbałość o szczegóły, a elementami zdradzającymi deepfake, są często niemal niewidoczne załamania krawędzi obrazu, nienaturalne deformacje krawędzi twarzy, nierealistyczne refleksy światła, czy niesynchroniczne mruganie lub ruch kącików ust, które podczas przeglądania treści w mediach społecznościowych, jest niemal niemożliwe do zauważenia[62].

Dodatkowym elementem utrudniającym prawidłową weryfikację treści deepfake jest możliwość ponownego wykorzystywania całości lub fragmentów już wcześniej opublikowanych i rozpowszechnianych materiałów, które mogą być wykorzystywane w niezmienionej formie, jak też poddawane dodatkowej obróbce w celu zmienienia ich wydźwięku oraz charakteru. Z uwagi na wykorzystywanie samouczących się algorytmów oraz dużą ilość dostępnych materiałów źródłowych, możliwe jest z pojedynczego materiału wejściowego otrzymanie innych, często odmiennych materiałów deepfake, które mimo wzajemnego wykluczania się, stanowią niezwykle wiarygodną treść i są jednocześnie elementem uzupełniającym określoną narrację. Takie rozwiązania mogą być wykorzystywane na potrzeby działań realizowanych w ramach prowadzonych operacji psychologicznych lub działań dezinformacyjnych, które ukierunkowane są na zaburzenie percepcji odbiorcy oraz uniemożliwienie mu podjęcia racjonalnych i zgodnych z jego wolą decyzji[63].

O ile w przypadku materiałów deepfake opartych o wideo lub fotografię, ich weryfikacja możliwa jest poprzez poszukiwanie niedoskonałości w wykonaniu, o tyle w przypadku materiałów obejmujących wyłącznie dźwięk,

[61] Zobacz: G. Pilarski, Cyberspace as a Tool…, s. 96.
[62] P. Galemba, op.cit., s. 226-227.
[63] P. Galemba, Operacje psychologiczne jako zagrożenie dla bezpieczeństwa państwa w XXI wieku, „Wiedza Obronna", 2025, nr 1 (290),s 196.

jego weryfikacja, bez zastosowania narzędzi, jest niemal niemożliwa, a w czasie wykorzystania jej na potrzeby przeprowadzenia rozmowy telefonicznej (np. w celu wyłudzenia danych, pieniędzy lub skłonienia do realizacji określonego działania), jest dla rozmówcy wręcz nierealna do wykonania, zwłaszcza wówczas, gdy nie zna rozmówcy osobiście i nie jest w stanie określić czy jego ton głosu, prędkość mówienia oraz sposób wypowiadanych słów jest odmienny od rzeczywistego.

Powyższe pozwala uznać, że dynamiczny rozwój technologii generowania treści syntetycznych sprawił, że identyfikacja i detekcja deepfake'ów staje się jednym z kluczowych problemów współczesnego świata cyfrowego. W miarę, jak algorytmy sztucznej inteligencji osiągają coraz wyższy poziom realizmu, zdolność odróżnienia materiałów autentycznych od zmanipulowanych ulega znacznemu osłabieniu. Proces wykrywania deepfake'ów nie jest już wyłącznie wyzwaniem technologicznym, lecz również społecznym, prawnym i poznawczym.

Podsumowanie

Przeprowadzone rozważania pozwalają stwierdzić, że technologia deepfake stanowi jedno z najbardziej złożonych i wielowymiarowych zjawisk współczesnej transformacji cyfrowej. Jej istota polega na wykorzystaniu zaawansowanych algorytmów sztucznej inteligencji, uczenia maszynowego i sieci neuronowych, do generowania lub modyfikowania treści audiowizualnych w sposób, który utrudnia odróżnienie fikcji od rzeczywistości. Autorzy wykazali, że deepfake nie jest jedynie narzędziem technologicznym, lecz zjawiskiem społeczno-kulturowym, które wpływa na sposób postrzegania informacji oraz wiarygodności przekazu[64].

Analiza podstaw definicyjnych pozwala stwierdzić, że deepfake wykracza poza tradycyjne formy manipulacji medialnej. W przeciwieństwie do klasycznych fotomontaży, technologia ta umożliwia tworzenie syntetycznych treści o wysokim stopniu realizmu, często generowanych w pełni automatycznie. Jej charakterystyczną cechą jest zdolność do naśladowania indywidualnych cech człowieka, takich jak mimika, głos, sposób mówienia czy gestykulacja, co znacząco zwiększa jej potencjał perswazyjny i manipulacyjny. Nie można kwestionować potrzeby uregulowania zjawiska deepfake'ów. Wymagana jest również intensywna współpraca najbardziej zaawansowanych podmiotów technologicznych świata nad poszukiwaniem skutecznych środków detekcji oraz unieszkodliwiania generowanych treści o potencjale szkodliwości[65].

Charakterystyka zjawiska deepfake ujawnia jego dualny charakter. Z jednej strony technologia ta znajduje zastosowanie w obszarach legalnych i społecznie pożądanych, takich jak przemysł filmowy, edukacja, rekonstrukcja historyczna czy dostępność cyfrowa. Z drugiej strony, jej powszechna dostępność i relatywnie niski próg wejścia sprzyjają nadużyciom, w tym dezinformacji, naruszeniom prywatności, oszustwom finansowym oraz destabilizacji debaty publicznej. Szczególnie niepokojące jest wykorzystanie deepfake'ów w kontekstach politycznych i społecznych, gdzie ich wpływ może wykraczać poza jednostkowe szkody, oddziałując na całe grupy społeczne. Wydaje się, że rozwój technologii generatywnych postępuje szybciej niż

[64] G. Pilarski, Cyberspace as a Tool of…, s. 107.
[65] D. Flisak, op.cit., s. 42

systemów detekcji, co prowadzi do asymetrii pomiędzy możliwościami tworzenia a zdolnością wykrywania manipulacji. W efekcie deepfake'y podważają tradycyjne mechanizmy weryfikacji informacji i zaufania do mediów.

W kontekście społecznym wykazano, że deepfake przyczynia się do erozji zaufania. Odbiorcy coraz częściej kwestionują autentyczność nawet prawdziwych materiałów, co sprzyja relatywizacji prawdy i wzmacnia zjawisko chaosu informacyjnego. Paradoksalnie, sama świadomość istnienia technologii deepfake może być wykorzystywana jako strategia obronna przez osoby rzeczywiście dopuszczające się nadużyć, które podważają autentyczność niekorzystnych dla nich dowodów.

Konkludując, technologia deepfake stanowi istotne wyzwanie dla współczesnego społeczeństwa informacyjnego. Jej znaczenie wykracza poza sferę technologiczną, obejmując obszary etyki, prawa, bezpieczeństwa oraz kultury medialnej. Skuteczna odpowiedź na zagrożenia związane z deepfake'ami wymaga podejścia interdyscyplinarnego, łączącego rozwój narzędzi detekcji, odpowiednie regulacje prawne oraz systematyczną edukację medialną. Tylko poprzez świadome i odpowiedzialne zarządzanie tą technologią możliwe jest wykorzystanie jej potencjału przy jednoczesnym ograniczaniu ryzyk, jakie niesie dla prawdy, zaufania społecznego i stabilności informacyjnej.

Jak trafnie wskazuje Ewelina Bartuzi-Trokielewicz „rozumiejąc techniki manipulacji, możemy budować społeczną odporność na fałszywe treści. Natomiast, żeby skutecznie przeciwdziałać tego typu zagrożeniom, potrzebujemy reakcji na bieżące incydenty, ale również potrzebna jest długofalowa strategia, która obejmie rozwój automatycznych metod detekcji oraz regulacje prawne, programy edukacyjne i standardy odpowiedzialnego korzystania z AI"[66].

Bibliografia

Bartuzi-Trokielewicz E., Techniki manipulacji w świecie deepfake'ów – jak sztuczna inteligencja zmienia oblicze dezinformacji, https://ai. gov.pl/aktualnosci/Techniki-manipulacji-w-%C5%9Bwiecie-deepfake'%C3%B3w.

Bateman J., Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios, Washington 2020.

Bernacka J., Problematyka prawna technologii deepfake – analiza legalności tworzenia i rozpowszechniania deepfake'ów po uchwaleniu AI Act, „Prawo i Więź" 2025, nr 5(58).

Bień-Węgłowska I., Deepfake w świetle aktu w sprawie sztucznej inteligencji, „Prawo i Więź", 2025, nr 5(58).

Bodanka O., Naruszenie wizerunku przy wykorzystaniu technologii deepfake – analiza prawna i praktyczna, „Opolskie Studia Administracyjno-Prawne" 2022, nr 1(20).

Chałubińska-Jentkiewicz K., Nowikowska M., Ochrona informacji w cyberprzestrzeni, Warszawa 2020.

Cymanow-Sosin K., Perswazja i manipulacja jako zjawiska kontekstowe na przykładzie reklam sektora usług instytucji rynku finansowego, „Naukowy Przegląd Dziennikarski" 2015, nr 2.

Ćwiertniak R., Ludzie i boty sztucznej inteligencji w nowoczesnych projektach, „Studia i Prace. Kolegium zarządzania i finansów" 2024, nr 199.

Dąbrowska I., Deepfake – nowy wymiar internetowej manipulacji, „Zarządzanie Mediami", 2020, nr 2 (8).

[66] E. Bartuzi-Trokielewicz, op.cit.

Flisak D., O niełatwym wytyczaniu granic rzeczywistości przez AI Act: zasada transparentności, deepfake, „Prawo Nowych Technologii" 2024, nr 1.
Galemba P., Deepfake jako zagrożenie i wyzwanie dla bezpieczeństwa informacyjnego państwa, „Cybersecurity and Law", 2025, nr 1(13).
Galemba P., Operacje psychologiczne jako zagrożenie dla bezpieczeństwa państwa w XXI wieku, „Wiedza Obronna", 2025, nr 1 (290).
Garcia Vera G., Deep fake – postęp technologiczny a prawo karne, „Acta Iuridica Resoviensia" 2024, nr 1(144).
Gibson A., Digital Humanities in the Deepfake Era, [in:] M.K. Gold, L.F. Klein (eds.), Debates in the Digital Humanities, University of Minnesota Press 2023.
Giles K., Hartmann K., Mustaffa M., The Role of deepfakes in mailing influence campaigns, Riga: NATO Strategic Communications Centre of Excellence 2019.
Himelein-Wachowiak M., Giorgi S., Devoto A., Rahman M. , Ungar L., Schwartz H.A., Epstein D.H., Leggio L., Curtis B., Bots and Misinformation Spread on Social Media: Implications for COVID-19, „Journal of Medical Internet Research" 2021, vol. 23, no. 5, doi: 10.2196/26933.
Jowett G.S., O'Donnell V., Propaganda & Persuasion, London 2018.
Kaczmarek K., Deepfake [w:] K. Chałubińska-Jentkiewicz (red.), Leksykon cyberbezpieczeństwa, Warszawa 2024.
Kiełpiński K. M., Deepfake jako narzędzie do przekazywania informacji fałszywej i domniemanej. Analiza prawnokarna i cybernetyczna, „Kwartalnik Krajowej Szkoły Sądownictwa i Prokuratury" 2023, nr 3(51).
Łabuz M., Deep fakes and the Artificial Intelligence Act - An important signal or a missed opportunity?, „Policy & Internet" 2024, nr 16, s. 11, https://doi.org/10/1002/poi3 .406.
Łukasik-Turecka A., Boty w służbie dezinformacji, „Polityka i Społeczeństwo" 2024, nr 4(22), DOI: 10.15584/polispol.2024.4.9.
Mularz K., Deepfake jako generator fikcyjnych treści i rzeczywistej nieufności, „Dydaktyka Informatyki" 2024, nr 19.
Nabiałek J., Wielka Brytania: Porno deepfake to przestępstwo od momentu utworzenia, https://forsal.pl/gospodarka/cyfryzacja /artykuly/9489216,wielka-brytania-porno-deepfake-to-przestepstwo-od-momentu-utworzenia .html.
NATO Strategic Communications Center of Excellence, Social Media as a Tool of Hybrid Warfare, Riga 2016, https://www.stratcomcoe.org/social-media-tool-hybrid-warfare.
NissenT.E., The Weaponization of Social Media, Royal Danish Defence College, Copenhagen 2015.
Nowikowska M., Deepfake w nowych mediach – zagrożenie dla tożsamości cyfrowej, „Roczniki Kulturoznawcze" 2025, vol. 17, nr 4.
Nowikowska M., Kozłowski M., Analiza wybranych zagrożeń dla tożsamości cyfrowej wynikających z postępu technologicznego, „Journal of Modern Science", 2025, nr 3(63).
Nowikowska M., PSYOPS as an elemnt of information warlfare, [in:] K. Chałubińska-Jentkiewicz, O. Evsyukova, Information, disinformation, cybersecurity, Toruń 2023.
Olech A. K., Lis, A., Wykorzystanie nowych technologii przez terrorystów na przykładzie dronów i deepfake'ów, „Wiedza Obronna", 2021, nr 2(275).
Ostrowski J.R., Shallowfakes, „The New Atlantis" 2023, no. 72, s. 96; https://www.jstor.org/ stable/ 27212358.
Piecuch A., Życie SMART – ma swoją cenę, „Dydaktyka Informatyki" 2020, nr 15.
Pilarski G, Cyberspace as a Tool of Contemporary Propaganda, „Safety&Defense" 2020, t. 6.
Pilarski G., Cyberprzestrzeń – relacje w wojnie hybrydowej, Warszawa 2020.
Pilarski G., Cyberspace as a Tool of Contemporary Propaganda, „Safety & Defense" 2020, tom 6 nr. 1, https://doi.org/10.37105/sd.72.

Pilarski G., Wirtualne relacje – szanse i zagrożenia wynikające z „życia" w cyberprzestrzeni, [w:] I. Horzela, A. Nowakowska-Krystman (red.), Zarządzanie relacjami wybrane aspekty, Warszawa 2015.

Rajtar O., Technologie deepfake w zakresie danych osobowych, „Rocznik Administracji Publicznej" 2025, nr 11(1).

Rozporządzenie 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 (Dz.Urz.UE L. z 12.7.2024).

Różyński P., Deepfake idzie na wojnę, „Rzeczpospolita", 22 marca 2022.

Sanocki A., Deepfakes, czyli postprawda objawiona, [w:] M. Wrzosek (red.), Zjawisko dezinformacji w dobie rewolucji cyfrowej. Państwo. Społeczeństwo. Polityka. Biznes, Warszawa 2019.

Sarowski Ł., Wpływ sztucznej inteligencji na infosferę – wybrane zagadnienia, „Tempus", 2024, nr 2.

Sewastianowicz M., Deepfake - ofiara realistycznej przeróbki może mieć problem z dochodzeniem swoich praw, https://www.prawo.pl/prawo/deepfake-a-prawo-karne-film-porno, 520138.html.

Sieńczyło-Chlabicz J., Przedmiot, podmiot i charakter prawa do wizerunku, „Przegląd Ustawodawstwa Gospodarczego" 2003, nr 8.

Smith H., Mansted K., Weaponised Deep Fakes: National Security and Democracy, Australia: Australian Strategic Policy Institute 2020.

Sobek J. P., Treści generowane przez sztuczna inteligencję w kontekście ochrony przed dezinformacją, „Studia bezpieczeństwa Narodowego" 2024, nr 33.

Spyra Z., Pyjas M., Wykorzystanie narzędzi deepfake w komunikacji merketingowej a społeczna odpowiedzialność biznesu z perspektywy pokolenia Z – wyniki badań eksploracyjnych [w:] H. Mruka, A. Sawicki (red.), Marketing. Koncepcje i doświadczenia, Pelpin 2024.

Szczepańska-Włoch J., (Nie)racjonalność jako strategia manipulacji. Wywiad polityczny w Wielkiej Brytanii, „Res Rhetorica" 2023, nr 10(2).

Thomas T.E. Nissen, The Weaponization of Social Media, Royal Danish Defence College, Copenhagen 2015.

Wasilewski L., Technologia deepfake wymaga regulacji – komputerowe przeróbki stają się zagrożeniem, https://www.prawo.pl/biznes/pierwsze-regulacje-prawne-technologii-deepfake,49 454.html.

Wasiuta O., Wasiuta S., Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość, „Annales Univeristatis Paedagogicae Cracoviensis. Studia de Securitate" 2019, nr 3(9).

Wojnicka E., Prawo do wizerunku w ustawodawstwie polskim, „Zeszyty Naukowe Uniwersytetu Jagiellońskiego. Prace z Wynalazczości i Ochrony Własności Intelektualnej" 1990, z. 56.

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

Zacher Sørensen M.M., Deepfake Face-Swap Animations and Affect, [w:] T. Tomoko (red.) Human Perception and Digital Information Technologies: Animation, the Body, and Affect, Bristol 2024.

Zalewski P., Nowe technologie, informacja, dezinformacja, fake news, deepfake, hejt. Wybrane mechanizmy manipulacji w przestrzeni cyfrowej w kontekście pandemii COVID-19, kryzysu granicznego w 2021 r. I wojny w Ukrainie, „Przegląd Policyjny" 2022, nr 2(146).

Ziobroń A., Deepfake a prawo karne. Uwagi de lege lata i de lege ferenda dotyczące fałszywej pornografii, „Studenckie Prace Prawnicze, Administratywistyczne i Ekonomiczne" 2021, nr 37.

Ziobroń A., Klonowanie głosu - wyzwaniem dla prawa karnego?: O przestępczości z użyciem sztucznej inteligencji, [w:] J. Piskorski, M. Błaszczak (red.), Nowe technologie: wyzwania i perspektywy dla prawa karnego, Łódź 2023.