ZARZĄDZANIE MEDIAMI
Tom 8(2)2020, s. 89–101
doi:10.4467/23540214ZM.20.024.11803
www.ejournals.eu/ZM

Ilona Dąbrowska

Uniwersytet Marii Curie-Skłodowskiej w Lublinie
ilona.anna.dabrowska@gmail.com
ORCID: 0000-0001-7451-2648

DEEPFAKE – NOWY WYMIAR INTERNETOWEJ MANIPULACJI

Abstract

DEEPFAKE – A NEW DIMENSION OF ONLINE LIES

The subject of consideration is the problem of falsification of audiovisual materials, referred to as deepfake. In this publication, I analyze literature on the problem, as well as draw the genesis and technological aspects of deepfake movies. Reflections on the issue discussed have been enriched with specific examples of films that have been published on the Internet to date. The article is an attempt to answer the question about the real threats associated with the existence of deepfakes and the possibilities of protection against them. An important aspect is the question about the security of every modern Internet user (in the context of data posted on social media), as well as the security of entire societies and nations. The publication highlights the threats associated with the spread of fake news and deepfakes in the online sphere in the context of off-line security.

Key words: deepfake, Internet, movies, recordings, social media, forgery, manipulation

Wstęp

Celem niniejszego artykułu jest analiza zjawiska określanego jako *deepfake*. Zaproponowana problematyka jest nie tylko aktualna, ale przede wszystkim niesłychanie istotna, szczególnie w kontekście dynamicznego rozwoju nowych technologii oraz niespotykanej wcześniej prędkości rozprzestrzeniania się treści (m.in. poprzez aplikacje mobilne w smartfonach). Przedmiot moich rozważań i badań stanowią rozpowszechnione w internecie materiały audiowizualne, które wykazują cechy materiałów celowo spreparowanych lub zostały wręcz opisane jako przejaw technologii *deepfake*. Problem badawczy stanowi pytanie o rozwój i rozpowszechnianie zarówno filmów typu *deepfake*, jak i wiedzy na temat możliwych negatywnych efektów

popularyzacji tego typu praktyk, oraz możliwości przeciwdziałania im. W celu realizacji badań zastosowano metodę analizy i krytyki piśmiennictwa, a także analizę materiałów audiowizualnych popularyzowanych w internecie. Dodatkową przesłanką przemawiającą za analizą niniejszego zjawiska jest istnienie luki informacyjnej: na gruncie polskim, jak i światowym brakuje literatury, która fachowo i kompleksowo podejmuje problem *deepfake'ów*.

Istota zjawiska

Deepfake to technologia wykorzystująca sztuczną inteligencję do tworzenia lub edytowania treści wideo albo obrazu w celu pokazania czegoś, co nigdy się nie wydarzyło[1]. *Deepfake* wideo jest tworzone przy użyciu dwóch konkurujących systemów AI – pierwszy to tzw. generator, drugi określany jest jako dyskryminator. Generator tworzy fałszywy klip wideo, a następnie dyskryminator ustala, czy klip jest prawdziwy, czy fałszywy. Za każdym razem, gdy dyskryminator dokładnie identyfikuje klip wideo jako fałszywy, daje on generatorowi wskazówkę, czego nie robić podczas tworzenia następnego klipu[2]. Termin „*deepfake*" łączy w sobie dwa zagadnienia: głębokie uczenie się oraz fałszerstwo. Jest to technika syntezy obrazów oparta na sztucznej inteligencji. Służy do łączenia i nakładania istniejących obrazów i filmów na obrazy źródłowe za pomocą specjalnej techniki uczenia maszynowego[3]. *Deepfake* to technologia obrazowania ludzi, która posługuje się sztuczną inteligencją do zmieniania obrazów ludzkich. Wykorzystuje algorytm określany jako GAN[4] – generatywne sieci przeciwstawne, aby nałożyć inne zdjęcie na zdjęcie źródłowe. Proces znany jest w cyfrowym świecie graficznym jako „nakładanie się"[5]. Co ciekawe, określenie „*deepfake*" nie miało funkcjonować jako termin dla filmów tego typu. W styczniu 2018 roku został wynaleziony i wprowadzony na rynek program o nazwie FakeApp. Podstawową funkcją tej aplikacji było umożliwienie ludziom tworzenia filmów z wymienionymi twarzami oraz udostępniania tych filmów. W celu wyprodukowania fałszywego wideo aplikacja wykorzystywała metodę głębokiego uczenia się.

Jeszcze całkiem niedawno specjaliści pochylali się nad pierwszymi materiałami typu *fake news*. Dziś, obok tego zjawiska, mamy do czynienia z nowym wymiarem

[1] N. Young, *DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media*, [niezależny wydawca], New York 2019, s. 14.

[2] M. Rouse, *Deepfake (deep fake AI)*, WhatIs.com, https://whatis.techtarget.com/definition/deepfake (dostęp: 29.10.2019).

[3] O. Schwartz, *You thought fake news was bad? Deep fakes are where truth goes to die*, „The Guardian", 12.11.2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth (dostęp: 13.11.2019).

[4] GAN to skrót od angielskiego terminu: *generative adversarial networks*.

[5] N. Young, dz. cyt., s. 14.

internetowej manipulacji – *deepfake'ami*. Termin „*fake news*" odnosi się do wiadomości medialnej, która jednocześnie nie jest ani prawdą, ani kłamstwem. Informacje tego typu opierają się na dezinformacji i często zawierają prawdziwe fragmenty[6]. *Słownik języka polskiego PWN* definiuje *fake news* jako: „zabieg manipulowania faktami, chętnie stosowany jest przez dziennikarzy, których celem, podczas przygotowywania publikacji, jest jak największe zainteresowanie tematem, a nie jego zgodność z rzeczywistością"[7]. *Fake news* ma za zadanie świadomie wprowadzić odbiorcę w błąd, na przykład po to, aby osiągać korzyści finansowe, polityczne czy propagandowe[8]. Podane definicje są bliskie sposobom opisywania *deepfake'ów*. Szczególnie ostatnia z nich, akcentująca cel tworzenia materiałów tego typu, wydaje się najbardziej trafna. *Deepfake* jest zjawiskiem stosunkowo świeżym, jednak w bardzo szybkim tempie ulega rozprzestrzenianiu. Niestety w parze z popularyzacją nagrań tego typu nie idzie świadomość i wiedza odbiorców filmów. Tym samym znajdujemy się w trudnym momencie, kiedy filmy stworzone za pomocą opisanych technologii oddziałują na odbiorców, którzy są zupełnie nieprzygotowani do percepcji *deepfake'ów*. Co gorsza – nie nadąża również ustawodawstwo.

Deepfake – geneza

Sama technologia zaczęła zyskiwać na popularności pod koniec 2017 roku, kiedy to na Reddicie[9] opublikowany został fałszywy film porno przedstawiający aktorkę Gal Gadot. Twórca filmu używał nicku *deepfakes* i właśnie od niego pochodzi używany dzisiaj termin. Wspomniany internauta w grudniu 2017 roku wykorzystał technologię głębokiego uczenia się, dodając twarze celebrytów aktorom w filmach pornograficznych. Od tego czasu w sieci pojawiło się wiele spreparowanych filmów z udziałem celebrytów i polityków – niektóre z nich mają być satyryczne, inne przedstawiają osoby publiczne w negatywnym świetle, a jeszcze inne stworzono, aby poprzeć konkretny punkt widzenia. Termin, który dotyczy zarówno technologii, jak i tworzonych przy jego użyciu filmów, jest zapowiedzią głębokiej nauki i podróbki.

Jeśli zastanowimy się nad ewolucją internetu oraz rozwojem rynku aplikacji mobilnych, nie powinniśmy być zdziwieni istnieniem *deepfake'ów*. Od kilku lat

[6] J. Gillin, *Fact-checking fake news reveals how hard it is to kill pervasive 'nasty weed' online*, PolitiFact, 27.01.2017, http://www.politifact.com/punditfact/article/2017/jan/27/fact-checking-fake-news-reveals-how-hard-it-kill-p/ (dostęp: 29.12.2019).

[7] Zob. www.sjp.pwn.pl (dostęp: 29.12.2019).

[8] M. Drzazga, *Cała prawda o fake news, czyli jak rozpoznać fałszywe wiadomości?*, https://www.legalniewsieci.pl/aktualnosci/cala-prawda-o-fake-news-czyli-jak-rozpoznacfalszywe-wiadomości (dostęp: 29.12.2019).

[9] Reddit to serwis internetowy przedstawiający linki do różnorodnych informacji, które ukazały się w sieci. Serwis jest głównie anglojęzyczny, chociaż jego interfejs jest tłumaczony na wiele języków (w tym polski).

w powszechnym użyciu są bowiem rozwiązania technologiczne, które umożliwiają zaawansowany retusz zdjęć. Każdy zainteresowany może samodzielnie „korygować" fotografie cyfrowe wedle potrzeb i gustu. Intuicyjny interfejs, błyskawiczne działanie – to tylko niektóre z cech aplikacji, których celem jest modyfikacja zdjęć. Narzędzia tego typu stały się rozwiązaniami pierwszej potrzeby dla osób biorących udział w wyścigu po internetowe uwielbienie (lajki). Zarówno osoby prywatne (choć prywatność w mediach społecznościowych stanowi ciekawe zagadnienie samo w sobie), jak i celebryci namiętnie sięgają po filtry i nakładki, które sprawiają, że internetowe *alter ego* jest w ich mniemaniu piękniejsze. Regularnie co jakiś czas jesteśmy świadkami afery związanej z nieudolnie wydłużonymi nogami czy zaokrąglonym parapetem na wysokości biustu. Kluczowe wydaje się nazewnictwo tego typu praktyk. Przyjęło się mówić o korygowaniu, poprawianiu, upiększaniu (jeden z filtrów ma nawet nazwę „upiększacz") itd. Tymczasem odpowiedniejszym określeniem wydaje się słowo: „fałszerstwo". Oczywiście w kontekście marketingowym aplikacje o nazwie „Oszukiwacze fotograficzne" nie miałyby szans. Dlatego też poszczególne filtry opakowano w piękne i pozytywnie nacechowane słowa, takie jak „nostalgia", „poranek" czy „miłość". Tymczasem tak bezkrytyczne podejście i chętne stosowanie filtrów skutkuje wieloma problemami o charakterze psychologicznym i społecznym. Z kolei moda na „najlepsze" (cokolwiek by to nie znaczyło) selfie zebrała już setki ofiar. Niestety także w dosłownym sensie. Temat śmierci podczas wykonywania tego typu zdjęć podjęli naukowcy z kilku uczelni medycznych w New Delhi. Opublikowany w czerwcu 2018 roku raport pt. *Selfie: dobrodziejstwo czy zmora*[10] prezentuje zaskakujące dane dotyczące ilości zgonów wśród fanów selfie. Analizie poddano 259 przypadków doniesień o śmierci osób robiących selfie, które wydarzyły w pomiędzy 2011 a 2017 rokiem na całym świecie. Wszystkie zostały potwierdzone. Dodatkowo naukowcy ustalili, że przyczyną śmierci w największej liczbie przypadków było utonięcie lub wypadek komunikacyjny (np. robienie selfie przed nadjeżdżającym pociągiem). Wysoko w klasyfikacji znalazły się także: upadek z wysokości, kontakt ze zwierzęciem, bronią czy porażenie prądem. Najwięcej przypadków utraty życia odnotowano kolejno w: Indiach, Rosji, Stanach Zjednoczonych i Pakistanie. Największy odsetek ofiar stanowiły osoby bardzo młode i młode (od 10. do 30. roku życia) – 85%. Skoro ludzie są w stanie ryzykować (a wręcz tracić) własne życie tylko po to, by zrobić zdjęcie, którym będą mogli pochwalić się w sieci, można wnioskować, że wszelkiego typu programy ułatwiające tworzenie spektakularnych zdjęć i filmów będą cieszyć się ogromną popularnością. Jak na razie użytkowników zajmują filtry i nakładki na zdjęcia, jednak skoro pojawia się analogiczny „ulepszacz" filmów – dlaczego by z niego nie korzystać?

Jak to możliwe, że komputer generuje coś, co nigdy nie istniało? A czyż nie do tego właśnie komputery zostały stworzone? Kto widział cyfry przechadzające się

[10] A. Bansal, Ch. Garg, A. Pakhare, S. Gupta, *Selfies: A boon or bane?*, „Journal of Family Medicine and Primary Care" 2018, vol. 7(4). Raport dostępny pod adresem internetowym: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131996/ (dostęp: 2.11.2019).

parkowymi uliczkami? Komputer liczy – zbiera i analizuje dane, przetwarza własne zasoby i wyrzuca wyniki. Jeśli zatem pewien system wyposażymy w zbiór fotografii, filmików i nagrań z naszym udziałem, za chwilę otrzymamy wyniki tego zbioru. I nikt nigdy nie powiedział, że będą to wyniki odzwierciedlające rzeczywistość. Skąd wzięły się multimedialne zbiory? W dużej mierze z naszego niefrasobliwego podejścia do mediów społecznościowych. W serwisach społecznościowych są miliardy zdjęć i nagrań, na których umieszczenie pozwoliliśmy.

Deepfake – próba typologii zjawiska

W czerwcu 2019 roku świat obiegło nagranie z udziałem Marka Zuckerberga[11] wygłaszającego krótką mowę, brzmiącą mniej więcej tak: „Wyobraź to sobie. Jeden człowiek z całkowitą kontrolą nad skradzionymi danymi miliardów ludzi. Wszystkie ich sekrety, ich życie, ich przyszłość". Nie byłoby w owym nagraniu nic dziwnego, gdyby nie fakt, że nie widzimy na nim prezesa Facebooka, lecz jego cyfrową replikę. Film wywołał spore zamieszanie i żywą dyskusję na temat aktualnych możliwości technologii cyfrowych. Nagranie szybko skategoryzowano jako przejaw *deepfake*. By przybliżyć sposób działania sztucznej inteligencji w tym obszarze, w sieci zaczęto publikować kolejne filmy z udziałem znanych osób, a dokładniej – z ich cyfrowymi odpowiednikami. Dużą popularnością cieszył się na przykład film z udziałem Kita Haringtona z popularnego serialu *Gra o tron* czy nagranie prezentujące prezydenta USA Baracka Obamę. O ile w pierwszym przypadku nagranie ma charakter ludyczny (Jon Snow grany Haringtona przeprasza za zakończenie serialu), o tyle przykład sięgania po wizerunki polityków może budzić uzasadnione obawy.

Zmanipulowane filmy i zdjęcia wykorzystujące sztuczną inteligencję mogą być wykorzystane zarówno do rozpowszechniania błędnych informacji, jak i do niszczenia reputacji konkretnych osób. W kwietniu 2018 roku BuzzFeed[12] zamieścił na swojej stronie film przedstawiający byłego prezydenta Baracka Obamę wypowiadającego słowa, które w rzeczywistości nigdy nie padły. Spreparowany cyfrowo Obama mówi między innymi: „Wkraczamy w erę, w której nasi wrogowie mogą zmusić kogokolwiek do wypowiedzenia dowolnych słów w dowolnym momencie. (...) To niebezpieczny czas. Musimy dużo bardziej uważać na internetowe treści"[13]. Twórcy serwisu chcieli w ten sposób zaakcentować zagrożenia, jakie może nieść ze sobą wykorzystanie technologii. W kolejnych miesiącach 2018 roku popularne stały się wideoklipy ze słynnych scen filmowych, które zostały cyfrowo zmienione, by

[11] Nagranie dostępne pod adresem: https://www.youtube.com/watch?v=cnUd0TpuoXI (dostęp: 30.10.2019).

[12] Serwis internetowy założony w 2006 roku przez Jonaha Peretti i Johna Johnsona. Witryna skupia się na rozrywce oraz wiadomościach.

[13] Film dostępny pod adresem: https://www.youtube.com/watch?v=cQ54GDm1eL0 (dostęp: 30.10.2019).

nałożyć na postacie twarz aktora Nicolasa Cage'a[14]. To przykład tej lżejszej i mniej szkodliwej strony problemu. Serwis społecznościowy Facebook znalazł się w ogniu krytyki po tym, jak odmówił usunięcia spowolnionego filmu z Nancy Pelosi, spikerką Izby Reprezentantów[15]. W tym przypadku film został spowolniony i zmontowany tak, by wydawało się, że Pelosi jest czymś odurzona. Mimo iż z technicznego punktu widzenia materiał ten nie stanowił czystego *deepfake'a*, wywołał dyskusję dotyczącą współczesnych możliwości spreparowania i rozpowszechniania filmów wideo. Niespełna rok od opisanych wydarzeń na rynku aplikacji mobilnych pojawił się program *DeepNude*. Funkcjonalność aplikacji polegała na tworzeniu realistycznych obrazów nagich kobiet za pomocą przesłanych zdjęć prawdziwych osób. Po krytycznych opiniach zrezygnowano z jej dystrybucji, jednak program został scrackowany i nadal jest dostępny w internecie.

Manipulacja cyfrowymi filmami i obrazami nie jest niczym nowym, ale nowe są już postępy w dziedzinie sztucznej inteligencji, łatwiejszy dostęp do narzędzi oraz skala, w jakiej można rozpowszechniać spreparowane materiały wideo. John Villasenor z Instytutu Brookingsa i Uniwersytetu Kalifornijskiego w Los Angeles twierdzi, że ostatnie dwa punkty są w dużej mierze powodem, dla którego *deepfaki* mogą budzić większe obawy niż pojawienie się w przeszłości innych narzędzi do edycji zdjęć i filmów. „Każdy jest teraz globalnym nadawcą – mówi Villasenor. – Myślę, że te dwie rzeczy tworzą fundamentalnie inne środowisko niż wtedy, gdy pojawił się Photoshop". Villasenor tłumaczy, że *deepfaki* tworzy się przy użyciu danych treningowych, tj. obrazów lub filmów wideo podmiotu, które służą do powstania trójwymiarowego modelu osoby. Ilość wymaganych danych może się różnić w zależności od używanego systemu i jakości podróbki, którą próbujesz stworzyć. Zdaniem Alego Farhadiego opracowanie przekonującego *deepfake'a* może wymagać tysięcy zdjęć i nagrań. Samsung opracował jednak system AI, który był w stanie wygenerować sfabrykowany klip wideo z użyciem zaledwie jednego zdjęcia[16].

W maju 2019 roku za sprawą sztucznej inteligencji „ożywiono" Mona Lisę z obrazu Leonarda da Vinci[17]. Stało się to możliwe dzięki opracowaniu system pozwalającego na stworzenie krótkich animacji na podstawie zaledwie jednego źródłowego kadru. Algorytmy kopiują mimikę, a następnie dzięki specjalnym znacznikom są w stanie animować nieruchome obrazy i zdjęcia twarzy. System określono mianem samouczącego się. Przykład Mona Lisy okazał się wyjątkowo trafiony

[14] Kompilacja filmów dostępna pod adresem: https://www.youtube.com/watch?v=BU9YAHigNx8 (dostęp: 30.10.2019).

[15] J. Waterson, *Facebook refuses to delete fake Pelosi video spread by Trump supporters*, „The Guardian", 24.05.2019, https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site (dostęp: 30.10.2019).

[16] L. Eadicicco, *There's a terrifying trend on the internet that could be used to ruin your reputation, and no one knows how to stop it*, „Business Insider", 10.07.2019, https://www.businessinsider.com/dangerous-deepfake-technology-spreading-cannot-be-stopped-2019-7?IR=T (dostęp: 30.10.2019).

[17] Film dostępny pod adresem: https://www.youtube.com/watch?v=P2uZF-5F1wI (dostęp: 3.12.2019).

– internauci bardzo szybko rozpowszechniali film, na którym słynna modelka robi miny, kręci głową i wygląda, jakby z kimś rozmawiała. Oczywiście nagranie służyło prezentacji najnowszej technologii, za którą tym razem odpowiedzialna była firma Samsung. Media na całym świecie prezentowały nagranie, zadając jednocześnie pytanie o kolejne możliwości nowych technologii. Twórcy zamieszania w opublikowanym on-line raporcie tłumaczą, że ich głównym wyzwaniem był problem syntezy fotorealistycznych spersonalizowanych obrazów głowy z zestawem znaków orientacyjnych twarzy, na których bazuje animacja modelu. Mechanizm ten ma mieć praktyczne zastosowanie do teleobecności, w tym do wideokonferencji i gier, a także do efektów specjalnych w branży filmowej[18].

W czerwcu 2019 roku w internecie pojawił się film, który wywołał spore poruszenie[19]. Na nagraniu widać humanoidalnego robota wykonującego serię militarnych ćwiczeń, by w ich końcowej fazie „zbuntować się" i zaatakować szkolących go ludzi. Podobnie jak w opisanych wyżej przypadkach, film wygląda całkowicie realnie. W ciągu zaledwie kilku dni wyświetlony został kilka milionów razy. Pojawiło się wiele komentarzy i artykułów – w większości mających na celu podkreślenie niebezpieczeństwa, jakie niesie ze sobą tworzenie robotów. Sami twórcy filmu zatytułowanego *Boston Dynamics: teraz do walki wchodzą nowe roboty* opisali produkcję słowami: „Dzięki dzisiejszej technologii okrucieństwo robotów stało się prawdziwym problemem". Co ważne, film w końcowej części demaskuje technologię, a zarazem kulisy stworzenia całej nieprawdziwej historii. Nie ma zatem wątpliwości co do sztucznej kreacji całego materiału. Okazało się jednak, że spora część internautów nie dotrwała do końca nagrania, kolejna część z kolei trafiła na wersję filmu pozbawioną tegoż wyjaśnienia (takie skrócone filmy powielano w serwisach społecznościowych). Materiał, który miał na celu poszerzenie wiedzy na temat nagrań typu *deepfake*, sam stał się tym, przed czym miał ostrzegać. Analogiczna sytuacja nastąpiła w przypadku filmu mającego na celu zwiększenie świadomości na temat porywania dzieci w Pakistanie[20]. Nagranie stworzone przez fundację Roshni Helpline prezentowało sfingowane porwanie dzieci bawiących się na ulicach Karaczi. Zarówno zakończenie, jak i napisy końcowe informują o celu powstania filmu. Niestety kilka lat po pierwotnym opublikowaniu nagrania za sprawą anonimowych internautów pojawiło się ono ponownie – już pozbawione pierwotnego zakończenia. W bardzo szybkim tempie przekazywano je kolejnym osobom (głównie poprzez popularną w Pakistanie aplikację WhatsUp). W krótkim czasie film trafił do tysięcy odbiorców, wzbudzając panikę, w wyniku której zginęło kilkudziesięciu

[18] I. Zakharo, A. Shysheya, E. Burkov, V. Lempitsky, *Few-Shot Adversarial Learning of Realistic Neural Talking Head Models*, raport dostępny pod adresem: https://arxiv.org/pdf/1905.08233.pdf (dostęp: 3.11.2019).

[19] Film dostępny pod adresem: https://www.youtube.com/watch?v=y3RIHnK0_NE (dostęp: 1.11.2019).

[20] Film dostępny pod adresem: https://www.youtube.com/watch?v=q5qdwyZJqzs (dostęp: 2.11.2019).

mężczyzn uznanych za członków przedstawionego na nagraniu „motocyklowego gangu porywaczy”[21]. Niebezpieczne konsekwencje miała także publikacja filmu z noworocznego orędzia prezydenta Gabonu, Alego Bongo Ondimby. Pojawił się on w sieci 1 stycznia 2019 roku, wywołując falę dezorientacji w kraju. W nietypowo krótkim, trzyminutowym przemówieniu oczy prezydenta wyglądały dość nienaturalnie, podobnie jak jego unieruchomione na krześle ciało. Po obejrzeniu filmu wiele osób w Gabonie myślało, że jest sfałszowany lub zmanipulowany. Kilka dni po upublicznieniu nagrania członkowie wojska Gabonu uznali, iż film stanowi wystarczający dowód na to, że Bongo nie nadaje się na prezydenta. W rezultacie zdecydowali się przeprowadzić zamach stanu, który jednak okazał się nieudany[22].

Przywołane przypadki *deepfake'ów* z pewnością nie są jedynymi istniejącymi w sieci materiałami audiowizualnymi tego typu. Stanowią one raczej pewien zestaw najbardziej wyrazistych przykładów nagrań mających na celu dezinformację czy manipulację. Jednak już na tej podstawie możliwe staje się dokonanie typologii analizowanego zjawiska. Czynnikiem, który uznałam za kluczowy w procesie porządkowania filmów typu *deepfake*, jest cel autora danego *deepfake'a*, tj. pierwotna intencja, powód, dla jakiego nagranie zostało stworzone. Zgodnie z tym założeniem wyróżniłam następujące cele: (1) rozrywka, (2) edukacja, (3) dezinformacja, (4) dyskredytacja.

Tabela 1. Typologia *deepfake'ów*

Cel *deepfake'a*	Główne cechy
Rozrywkowy	Charakter ludyczny. Często odwołuje się do popkultury, bohaterami są zwykle celebryci, aktorzy, osoby anonimowe i fikcyjne.
Edukacyjny	Ma na celu edukowanie odbiorcy. Często przy wykorzystaniu wizerunków osób znanych, w tym nieżyjących.
Dezinformacyjny	Wywołuje dezinformację, szum medialny, niepokój społeczny. Dotyczy zarówno osób publicznych, jak i prywatnych.
Dyskredytacyjny	Ma na celu osłabienie pozycji danej osoby, grupy, organizacji czy marki. Najczęściej obiektem tego typu *deepfake'ów* są politycy.

Źródło: opracowanie własne.

Rozważając genezę powstawania *deepfake'ów*, należy szczególnie podkreślić pierwotną intencję tworzenia nagrania. Filmy tego typu są bowiem podatne na mutacje – kolejne wersje, przeróbki, które nadają każdorazowo nowy sens, a także mogą wpływać na zmianę celu publikowania. W tym kontekście *deepfake* jest

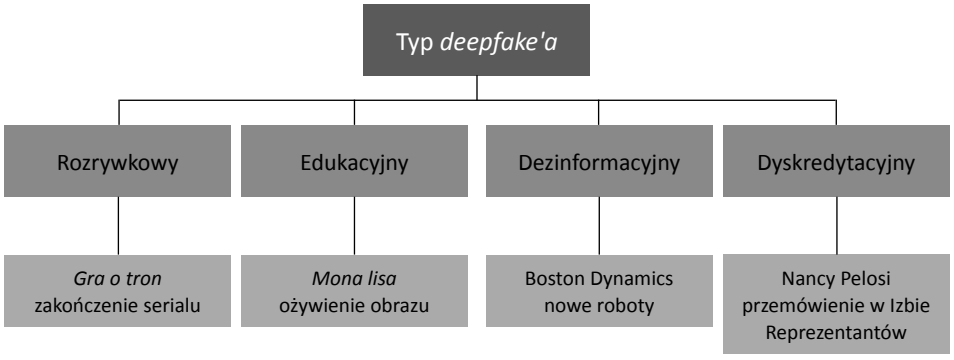
[21] B. Kicior, *To fake video doprowadziło do śmierci co najmniej 20 osób*, Menway, 21.11.2018, https://menway.interia.pl/meskie-tematy/news-to-fake-video-doprowadzilo-do-smierci-co-najmniej-20-osob,nId,2688001 (dostęp: 2.11.2019).

[22] T. Joplin, *A Military Coup in Gabon Inspired by a Potential Deepfake Video Is Our Political Future*, Albawaba News, 8.05.2019, https://www.albawaba.com/news/military-coup-gabon-inspired-potential-deepfake-video-our-political-future-1284760 (dostęp: 29.12.2019).

podobny do internetowego memu. Raz opublikowany, w dalszym ciągu może funkcjonować w sieci, zarówno w wersji oryginalnej, jak i kolejnych „odmianach". Jednak każda następna modyfikacja nie pozostaje bez wpływu na sens całości, a co za tym idzie – realizowany czy też osiągany cel.

Posługując się zaproponowanym podziałem, można dokonać kwalifikacji omawianych powyżej *deepfake'ów*. Dobrym przykładem *deepfake'a* o charakterze rozrywkowym jest film przedstawiający zakończenie serialu *Gra o tron*. Nagranie bardzo szybko rozpowszechniło się w internecie, ale przyniosło jedynie pozytywne bądź neutralne społecznie skutki. Podobny ładunek niosą ze sobą *deepfaki* edukacyjne. Mogą one służyć zarówno tłumaczeniu samej technologii „podrabiania" filmów (tzw. *metadeepfaki*), jak i edukować w dowolnie wybranym obszarze oraz zakresie (np. poprzez tworzenie wykładów z udziałem nieżyjących już autorytetów z danej dziedziny). O ile w tych dwóch przypadkach można mówić o pozytywnym wykorzystaniu zdobyczy technologicznych (choć nie brakuje także głosów krytykujących „ożywianie" postaci na potrzeby filmów), o tyle pozostałe dwie kategorie, tj. *deepfaki* dezinformacyjne i dyskredytacyjne, są niebezpieczne zarówno dla samych osób będących bezpośrednimi ofiarami filmów, jak i dla całych mas odbierających dany kontent. Poniższy wykres prezentuje przykładowy podział wybranych *deepfake'ów* zgodnie z zaproponowaną typologią.

Wykres 1. Podział wybranych *deepfake'ów* ze względu na ich cel



Źródło: opracowanie własne.

Zagrożenia i przeciwdziałanie

Choć część *deepfake'owych* filmów służy wyłącznie rozrywce, jasne jest, że technologia ta może posłużyć konkretnym intrygom. Nietrudno wyobrazić sobie hipotetyczną sytuację, kiedy na kilka dni przed wyborami w sieci pojawia się nagranie polityka, który

okazuje się antysemitą czy homofobem. Wygląda na pijanego, obiecuje absurdalne ustawy, grozi swojej rodzinie. Spreparowane za pomocą sztucznej inteligencji nagranie może zmienić wynik wyborów, zdyskredytować daną osobę, pozbawić ją godności.

Świadomość na temat nowinek technologicznych, a co za tym idzie – możliwości fałszowania materiałów audiowizualnych stopniowo ulega zwiększeniu. Dziś zastanawiamy się nad sposobami zabezpieczenia przed kolejnymi spreparowanymi nagraniami wideo. Oczywiste jest, że raz dana technologia nie zostanie zapomniana. Trudno też wyobrazić sobie wprowadzenie restrykcji dotyczących tego typu praktyk w sieci. Warto zastanowić się nad pytaniem o to, czy sama technologia jest niebezpieczna, czy raczej intencje i zachowania użytkowników sieci. Idąc dalej: czy zakaz tworzenia programów wykorzystujących technologię tego typu rozwiąże problem? W wywiadzie dla Business Insider Maneesh Agrawala, profesor informatyki z Forest Baskett i dyrektor Brown Institute for Media Innovation na Uniwersytecie Stanforda, zaznacza, że to nie technologia, lecz sposób jej wykorzystania stanowi największy problem. Tym samym wyeliminowanie *deepfake'ów* nie usunie źródła problemu. Zwraca on uwagę na to, że błędne informacje mogą być prezentowane także w przypadku, kiedy dany film jest w stu procentach prawdziwy. Nie powinniśmy zatem skupiać się na samej technologii, a raczej na praktykach dezinformacyjnych[23]. Zasadne więc staje się pytanie o to, co należy zrobić, aby zapobiec wykorzystywaniu *deepfake'ów*. Problem ten pojawił się już na stopniu państwowym w Stanach Zjednoczonych. W grudniu 2018 roku senator Ben Sasse zaproponował ustawę o nazwie *Malicious Deep Fake Prohibition Act of 2018*[24], której głównym celem jest zakazanie preparowania nagrań audiowizualnych. Projekt spotkał się jednak z krytyką, głównie ze względu na luki. Uwagi zgłoszono również do definicji zjawiska, brzmiącej w projekcie tak: „*Deepfake* oznacza zapis audiowizualny utworzony lub zmieniony w taki sposób, że zapis ten fałszywie wydaje się rozsądnemu obserwatorowi autentycznym zapisem faktycznej wypowiedzi lub zachowania danej osoby". Według krytyków tego typu zapis ogranicza zakres ustawy do zakazania tylko tych *deepfake'ów*, które nie są wyraźnie oznaczone jako takie[25]. Kolejne działania w tej sprawie zaproponowała reprezentantka demokratów Yvette Clarke. Jej pomysł polegał na wprowadzeniu nakazu umieszczania znaku wodnego na nagraniach tego typu. Proponowany projekt ustawy o nazwie: *Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019*[26] miałby też nakładać kary za nieprzestrzeganie nowych zasad.

[23] L. Eadicicco, dz. cyt.

[24] Projekt ustawy dostępny pod adresem: https://www.congress.gov/bill/115th-congress/senate-bill/3805/text (dostęp: 30.10.2019).

[25] S. Lahoti, *Lawmakers introduce new Consumer privacy bill and Malicious Deep Fake Prohibition Act to support consumer privacy and battle deepfakes*, Pact, 5.02.2019, https://hub.packtpub.com/lawmakers-introduce-new-consumer-privacy-bill-and-malicious-deep-fake-prohibition-act-to-support-consumer-privacy-and-battle-deepfakes/ (dostęp: 2.11.2019).

[26] Dokument dostępny pod adresem: https://www.congress.gov/bill/116th-congress/house-bill/3230 (dostęp: 2.11.2019).

We wrześniu 2019 roku Facebook i Microsoft wydali oświadczenie, że wspólnie występują przeciw fałszywym treściom w technologii *deepfake*. Jednym z pomysłów na walkę z fałszywymi treściami ma być Deepfake Detection Challenge – program wspierający badania nad metodami wykrywania fałszywych treści[27]. Niespełna miesiąc później do programu dołączył kolejny gigant – firma Amazon. W międzyczasie Google zdecydowało się na stworzenie trzech tysięcy własnych filmów *deepfake* w celu uczenia systemu odróżniania fałszywek od oryginałów[28]. Niestety dużym problemem cały czas pozostaje grunt prawny. W telewizyjnym wywiadzie dla CBS (nagranym w czerwcu 2019 roku) prezes serwisu Instagram otwarcie przyznał, że jego firma nie posiada jeszcze polityki dotyczącej *deepfake'ów*, lecz oczywiście nad tym pracuje. Problematyczne jest jednak już samo poprawne definiowanie tego zjawiska. Kolejne wyzwanie stanowi określenie zasad, według których należałoby działać, tak aby nie łamać istniejących regulaminów oraz prawa[29]. Co ciekawe, Instagram nie usunął fałszywego wideo z udziałem Marka Zuckerberga właśnie w obawie o pogwałcenie własnych zasad. Okazuje się zatem, że nie nadążają nie tylko organy prawa, ale także firmy zajmujące się nowymi technologiami i mediami społecznościowymi. W przypadku Unii Europejskiej możemy mówić o podejmowaniu tego tematu – przewija się on w dyskusjach i prelekcjach. Jednak w dalszym ciągu nie opublikowano dokumentu regulującego i wspierającego walkę z tym niebezpiecznym zjawiskiem. UE przyjęła strategię straszenia wprowadzeniem regulacji – w ten sposób chce zmusić firmy do stworzenia własnych zasad i działań[30].

Zakończenie

Technologia służąca fałszowaniu nagrań wideo może być wykorzystywana w rozmaity sposób i do różnych celów. Jednym z pól objętych tym procederem może stać się arena polityczna. Nietrudno wyobrazić sobie wojnę polityczną, w której bronią będzie sztuczna inteligencja generująca nagrania dowolnej treści. Nietrudno także wyobrazić sobie nowy wymiar manipulacji oraz dezinformacji. Ogromną potrzebą, a zarazem dużym wyzwaniem jest zatem wprowadzenie przepisów, które miałyby na celu rozwiązanie problemu *deepfake'ów* skutecznie i jednocześnie bez naruszania

[27] Szczegółowy opis i regulamin programu dostępny pod adresem: https://deepfake-detection-challenge.ai/ (dostęp: 3.11.2019).

[28] K. Hao, *Google has released a giant database of deepfakes to help fight deepfakes*, MIT Technology Review, 25.09.2019, https://www.technologyreview.com/f/614426/google-has-released-a-giant-database-of-deepfakes-to-help-fight-deepfakes/ (dostęp: 3.11.2019).

[29] Cały wywiad dostępny pod adresem: https://www.cbsnews.com/news/adam-mosseri-interview-instagram-head-on-deepfakes-exclusive/ (dostęp: 3.11.2019).

[30] T. Chivers, *What do we do about deepfake video?*, „The Guardian", 23.06.2019, https://www.theguardian.com/technology/2019/jun/23/what-do-we-do-about-deepfake-video-ai-facebook (dostęp: 3.11.2019).

wolności słowa czy aktualnych aktów prawnych. Niewątpliwie najbardziej skutecznym narzędziem do walki z tego typu nagraniami wydaje się dzisiaj stworzenie technologii deszyfrującej – tj. takiej, która rozpozna, że dany film został spreparowany. Kluczowe jest tu jednak słowo: dzisiaj. Technologia rozwija się bowiem w tak szybkim tempie, że wprowadzone rozwiązanie za chwilę może się okazać bezużyteczne. Zespołom pracującym nad danym problemem trudno jest wyprzedzić tych, którzy wyrządzają szkody. Pewne jest także to, że dalszego rozwoju technologii służącej *deepfake'om* nikt nie jest w stanie zatrzymać.

Bibliografia

Bansal A., Garg Ch., Pakhare A., Gupta S., *Selfies: A boon or bane?*, „Journal of Family Medicine and Primary Care" 2018, vol. 7(4). Raport dostępny pod adresem: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131996/ (dostęp: 2.11.2019).

Chivers T., *What do we do about deepfake video?*, „The Guardian", 23.06.2019, https://www.theguardian.com/technology/2019/jun/23/what-do-we-do-about-deepfake-video-ai-facebook (dostęp: 2.11.2019).

Drzazga M., *Cała prawda o fake news, czyli jak rozpoznać fałszywe wiadomości?*, https://www.legalniewsieci.pl/aktualnosci/cala-prawda-o-fake-news-czyli-jak-rozpoznacfalszywe--wiadomości (dostęp: 29.12.2019).

Eadicicco L., *There's a terrifying trend on the internet that could be used to ruin your reputation, and no one knows how to stop it*, „Business Insider", 10.07.2019, https://www.businessinsider.com/dangerous-deepfake-technology-spreading-cannot-be-stopped-2019-7?IR=T (dostęp: 30.10.2019).

Gillin J., *Fact-checking fake news reveals how hard it is to kill pervasive 'nasty weed' online*, PolitiFact, 27.01.2017, http://www.politifact.com/punditfact/article/2017/jan/27/fact-checking-fake-news-reveals-how-hard-it-kill-p/ (dostęp: 29.12.2019).

Hao K., *Google has released a giant database of deepfakes to help fight deepfakes*, MIT Technology Review, 25.09.2019, https://www.technologyreview.com/f/614426/google-has-released-a-giant-database-of-deepfakes-to-help-fight-deepfakes/ (dostęp: 3.11.2019).

Joplin T., *A Military Coup in Gabon Inspired by a Potential Deepfake Video Is Our Political Future*, Albawaba News, 8.05.2019, https://www.albawaba.com/news/military-coup-gabon-inspired-potential-deepfake-video-our-political-future-1284760 (dostęp: 29.12.2019).

Kicior B., *To fake video doprowadziło do śmierci co najmniej 20 osób*, Menway, 21.11.2018, https://menway.interia.pl/meskie-tematy/news-to-fake-video-doprowadzilo-do-smierci--co-najmniej-20-osob,nId,2688001 (dostęp: 2.11.2019).

Lahoti S., *Lawmakers introduce new Consumer privacy bill and Malicious Deep Fake Prohibition Act to support consumer privacy and battle deepfakes*, Pact, 5.02.2019, https://hub.packtpub.com/lawmakers-introduce-new-consumer-privacy-bill-and-malicious-deep-fake-prohibition-act-to-support-consumer-privacy-and-battle-deepfakes/ (dostęp: 2.11.2019).

Rouse M., *Deepfake (deep fake AI)*, WhatIs.com, https://whatis.techtarget.com/definition/deepfake (odczyt 29.10.2019).

Schwartz O., *You thought fake news was bad? Deep fakes are where truth goes to die*, „The Guardian", 12.11.2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth (dostęp: 13.11.2019).

Waterson J., *Facebook refuses to delete fake Pelosi video spread by Trump supporters*, „The Guardian", 24.05.2019, https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site (dostęp: 30.10.2019).

Young N., *DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media*, [niezależny wydawca], New York 2019.

Zakharo I., Shysheya A., Burkov E., Lempitsky V., *Few-Shot Adversarial Learning of Realistic Neural Talking Head Models*, raport dostępny pod adresem: https://arxiv.org/pdf/1905.08233.pdf (dostęp: 3.11.2019).