Klaudia Proszowska*

PERCEPCJA WYKORZYSTANIA TECHNOLOGII DEEPFAKE W MATERIAŁACH WIDEO WŚRÓD STUDENTÓW

Niniejszy artykuł skupia się na opiniach dotyczących wykorzystania technologii deepfake w materiałach wideo uznawanych za szczególnie niebezpieczne w kontekście rozpowszechniania fałszywych treści. Badanie pomogło określić, jak studenci rozumieją i oceniają deepfaki oraz jak ich opinia dotycząca szkodliwości (na poziomie społeczeństwa, osoby przedstawionej w materiale oraz osoby oglądającej) i konieczności oznaczenia fałszywości materiału zależy od wyróżnionych pięciu kategorii deepfake'ów: projektu artystycznego, satyry, materiału propagandowego, fake newsa oraz materiału pornograficznego. Sprawdzono również, czy te opinie są zależne od stopnia znajomości technologii. Badanie pokazało ogólnie negatywny stosunek do deepfake'ów, a za najbardziej szkodliwe na wszystkich płaszczyznach uznaje się materiały pornograficzne, projekty artystyczne z kolei – za najmniej szkodliwe. Pomimo nieznacznych różnic między kategoriami materiałów deepfake respondenci opowiadali się zwykle za ich czytelnym oznaczaniem. Znajomość technologii raczej nie ma bezpośredniego wpływu na postawy ankietowanych, chociaż osoby deklarujące jej lepszą znajomość uważają deepfaki za nieco bardziej przyszłościowe i interesujące, a materiały pornograficzne za mniej szkodliwe.

THE PERCEPTION OF DEEPFAKE TECHNOLOGY IN VIDEO MATERIALS AMONG UNIVERSITY STUDENTS

This research focuses on public opinion regarding the use of deepfake technology in video materials; materials considered particularly dangerous in terms of fake content dissemination. It studies how students understand and evaluate deepfakes and their opinion regarding harmfulness of deepfakes at three levels: at the level of society, for the person presented in the material and for the viewer. The research investigates also whether deepfakes shall be clearly marked as fake. Deepfakes were considered in five categories: as an art project, a satire, propaganda material, fake news or a pornographic material. The link to the user's proximity to technology was also explored. Generally, the attitude towards deepfakes was rather negative, with pornographic materials seen as the most harmful, and artistic projects – as the least. Regardless of the category, most respondents were in favor of clear labels on deepfake materials. Proximity to technology does not have a direct impact on the declared attitudes, although people closest to technology consider pornographic deepfake materials as less harmful, and general deepfake technology to be slightly more interesting and relevant in the future.

* AGH Akademia Górniczo-Hutnicza, Wydział Humanistyczny.

1. WSTĘP

Stworzyć wizerunek osoby, która nigdy nie istniała, wstawić twarz słynnego aktora do filmów, w których nigdy nie zdecydowałby się wystąpić, usłyszeć głos dowolnej osoby wypowiadającej nieprawdopodobne kwestie – to tylko niektóre z możliwości technologii deepfake, która opierając się na uczeniu maszynowym i dowolnym, wybranym przez twórcę materiale oraz bazie danych, tworzy materiały audiowizualne i tekstowe o każdej tematyce.

Technologia ta zyskała na popularności w 2017 roku, kiedy to internet podbiły sfałszowane filmy pornograficzne wykorzystujące wizerunki celebrytek. Od tego czasu deepfake jest przedmiotem wielu rozważań w kontekście stosowania tej technologii, a zwłaszcza płynących z niej zagrożeń i korzyści. Pomimo skali zagrożeń stosunkowo niewiele państw reguluje prawnie publikację deepfake'ów, a tym samym możliwą ochronę potencjalnych ofiar takich materiałów. Każda osoba publikująca treści zawierające własny wizerunek może stać się bohaterem deepfake'u, a zależnie od charakteru stworzonego materiału – również jego ofiarą. W momencie pisania tego artykułu prawo nie chroni osób i ich wizerunków wykorzystywanych do tworzenia treści deepfake. Dlatego też rozpowszechnienie i forma materiałów deepfake zależą głównie od internautów – ich potencjalnych konsumentów, od ich opinii i przyzwolenia na powstawanie i rozpowszechnianie różnych rodzajów materiałów deepfake.

Z tego powodu przedmiotem niniejszych badań[1] jest opinia studentów na temat materiałów wideo stworzonych za pomocą technologii deepfake. W porównaniu z innymi rodzajami treści deepfake materiały wideo niosą ze sobą podwyższony stopień ryzyka, zwłaszcza w czasach postępującej dezinformacji. Studenci zaś należą do jednej z grup wiekowych, które publikują najwięcej treści w internecie, a ze względu na swój wiek mogą być traktowani jako osoby kształtujące przekonania dotyczące tego, jakie treści są odpowiednie do publikowania, a jakie nie.

Na potrzeby tego badania stworzono pięć kategorii wideo będących deepfake'ami: projekt artystyczny, satyra, materiał propagandowy, fake news oraz materiał pornograficzny. Badanie ma pokazać, czy postrzegana szkodliwość deepfake'ów w trzech płaszczyznach (dla społeczeństwa, osoby przedstawionej w materiale oraz osoby oglądającej materiał) różni się w zależności od treści materiału, czy może jednak ich szkodliwość (lub jej brak) jest dla nich wszystkich jednakowa. W badaniu poruszana jest również kwestia istotności oznaczania deepfake'ów jako materiałów fałszywych i tu podobnie – ma ono zweryfikować, czy konieczność oznaczania takich materiałów jest taka sama dla wszystkich wyróżnionych kategorii. Badanie ma też wykazać, czy opinie respondentów zależą od ich płci, studiowanego kierunku (specyfiki oraz stopnia) oraz ogólnie pojętego kontaktu z technologią.

2. PRZEGLĄD LITERATURY

Zgodnie z definicją, którą sformułowali Giles i inni [1], deepfake (DF) to jakikolwiek materiał cyfrowy, który został wytworzony z zastosowaniem metod uczenia maszynowego, gdzie szczególnie wykorzystuje się generatywne sieci antagonistyczne (GAN). Za początek DF

[1] Badania zostały przeprowadzone w ramach pracy licencjackiej autorki, a sam artykuł powstał na jej podstawie.

w internecie uważa się publikację sfałszowanych filmów pornograficznych z wykorzystaniem wizerunków celebrytek (m.in. Gal Gadot) w serwisie Reddit przez anonimowego użytkownika [2]. Od tamtego czasu DF znacznie się rozpowszechniły, a ich tematyka stała się bardziej różnorodna – od nieszkodliwych przeróbek popularnych filmów [3], do narzędzi służących dezinformacji [4]. Bazując na sposobie stworzenia DF, można wyróżnić kilka ich rodzajów, jednak wytworzony DF może przyjąć jedną z następujących postaci: graficzny, audio, wideo lub tekstowy [5].

W kontekście popularyzacji DF tematem wielokrotnie rozważanym były różne możliwości ich zastosowań i to, w jaki sposób wpłyną one na jednostki i całe społeczeństwo [5–7]. Tak samo jak w przypadku wielu technologii wśród zastosowań DF znajdą się te pozytywne, ale i te, które niosą za sobą zagrożenia. Beneficjentami technologii DF mają być sfery edukacji, szeroko rozumianej sztuki i sposobów ludzkiej ekspresji [6]. Wśród negatywnych skutków szczególnie widoczny jest podział na zagrożenia dla jednostek (m.in. wyłudzenia) i społeczeństwa ogólnie (m.in. pogorszenie stosunków międzynarodowych, ale też i tych wewnątrz państwa) [5, 6].

Jak zauważa Harris [7], w momencie rozpowszechnienia się technologii DF, kiedy głównym sposobem jej wykorzystania były filmy pornograficzne, prawo nie chroniło jej ofiar, jednak od tamtego czasu w kontekście prawnym dotyczących DF wiele się zmieniło – mimo tego, że przykładów aktów prawnych regulujących DF wciąż nie jest zbyt dużo, są kraje oraz związki, które zdają sobie sprawę z zagrożeń wynikających z tych materiałów i pracują nad regulacjami [5].

Dotychczasowe badania dotyczące różnych aspektów DF pokazują, jak istotna i aktualna jest problematyka związana z DF w formie audiowizualnej, w których przypadku negatywne skutki są widoczne nie tylko w sferze rozważań. Dobber i inni [8] badali wpływ DF w formie wideo na opinię małej społeczności o polityku i partii, którą reprezentuje. Poparcie dla samego polityka wśród badanych spadło, podczas gdy dla partii, którą reprezentuje – nie zmieniło się [8]. To pokazuje, jak istotne jest rozróżnienie skutków konkretnego materiału dla jednostki, której wizerunek jest wykorzystywany, oraz dla społeczności – wpływ danego materiału DF może się różnić na tych dwóch płaszczyznach.

Kolejnym z często badanych aspektów DF w formie wideo jest stopień rozpoznania nieprawdziwości materiału. Badania Vaccariego oraz Chadwicka [9] w tym nurcie pokazały, że szczególnie istotny jest odsetek osób, które są niepewne w ocenie prawdziwości materiału. Twórcy badania wnioskują, że sama ta niepewność jest negatywnym skutkiem DF – zmniejszają one zaufanie do mediów i informacji w nich przedstawionych [9].

Warto zwrócić uwagę na to, że jednym z najczęściej wymienianych skutków rozwoju i popularyzacji technologii DF jest właśnie dezinformacja [6]. Bazując na badaniach przeprowadzonych przez Sundar i innych [10], można stwierdzić, że to materiały wideo mogą najbardziej przyczynić się do rozpowszechniania dezinformacji. Z tego powodu to DF w formie wideo mogą być szczególnie istotne w badaniu szkodliwości materiałów stworzonych za pomocą tej technologii. DF mogą być jednak kojarzone także z humorystycznymi filmikami, gdzie np. niskiej jakości wizerunki polityków wypowiadają różne zabawne kwestie. Te skrajne, ale obecne w sieci przypadki pokazują, że każdy użytkownik internetu może mieć całkowicie różne doświadczenie z DF, a tym samym – opinie na ich temat.

Przytoczone badania pokazują, że technologia DF jest ogromnie złożonym zjawiskiem, a także że problematyka DF jest niezwykle istotna w kontekście przyszłości mediów w internecie oraz ich percepcji. Dlatego też przedmiotem niniejszego badania są opinie studentów na temat DF w formie wideo, ale również to, w jaki sposób studenci rozumieją sam termin DF. Ze względu na różny charakter i tematykę prezentowanych materiałów wideo DF badanie pokaże, czy opinia respondentów na temat technologii DF zmienia się ze względu na tę tematykę, a jeśli tak, to w jaki sposób. Oprócz tego studenci zostaną zapytani o to, DF o jakiej treści powinny być oznaczane przez ich autorów jako fałszywe.

3. OPIS BADAŃ

3.1. PODSTAWOWE POJĘCIA ANALIZY

W niniejszym badaniu materiały DF będą rozumiane jako materiały audiowizualne wytworzone za pomocą technik uczenia maszynowego pokazujące sytuacje, które nigdy się nie wydarzyły, w tym nieistniejące wypowiedzi osób, których wizerunek jest wykorzystany w materiale.

Materiały DF w formie wideo zostały podzielone na pięć kategorii ze względu na ich charakter, treść oraz cel ich stworzenia (tab. 1). Kategorie w poniższej klasyfikacji są nierozłączne, co sprawia, że jeden konkretny materiał może przynależeć do więcej niż jednej kategorii. Podział jest autorski i powstał na potrzeby analityczne, by móc lepiej zakwalifikować dane, szczególnie w przypadku badania zmian opinii studentów wobec wyszczególnionych rodzajów DF.

Tabela 1. Podział deepfake'ów na kategorie oraz ich definicje

Kategoria	Definicja
Projekt artystyczny	Materiał, który powstał przez wykorzystanie wizerunku osoby publicznej lub aktora i nie ośmiesza osoby przedstawionej
Satyra	Materiał, który powstał przez wykorzystanie wizerunku osoby publicznej lub aktora i ośmiesza osobę przedstawioną
Materiał propagandowy	Materiał propagujący konkretne wartości, które mają wpłynąć na myślenie odbiorców, ale jego celem nie jest propagowanie fałszywych informacji
Fake news	Materiał mający za cel rozpropagowanie fałszywych informacji i/lub treści
Materiał pornograficzny	Materiał, który powstał przez sfabrykowanie filmu pornograficznego z wykorzystaniem czyjegoś wizerunku bez zgody tej osoby

Podstawą do stworzenia tego podziału są m.in. rozważania Chesneya i Citron [6] na temat potencjalnego zastosowania DF, gdzie możliwości wykorzystania takiego typu materiałów podzielono ze względu na skutki, jakie za sobą niosą – pozytywne i negatywne. Warto zwrócić uwagę na to, że jest dużo więcej zauważalnych zagrożeń niż korzyści.

Ta dysproporcja widoczna jest również w opracowanej klasyfikacji – trzy z wyszczególnionych kategorii charakteryzuje widocznie szkodliwy cel wytworzonych treści. Mowa tutaj o satyrze (ośmieszającej osobę przedstawioną), fake newsach (rozprzestrzenianie fałszywych informacji) oraz filmie pornograficznym (wykorzystanie czyjegoś wizerunku do stworzenia materiału pornograficznego). W przypadku tego ostatniego trudno określić konkretny cel przyświecający autorowi, jednak bazując na rozważaniach dotyczących DF w tej sferze, można uznać, że celem często jest upokorzenie osoby przedstawionej czy wymuszenie jakichś zachowań tej osoby [7]. Pozostałe kategorie – projekt artystyczny i materiał propagandowy – mimo pozytywnego zabarwienia dalej mogą nieść za sobą negatywne skutki, które nie były przewidziane przez autora DF, dlatego można uznać ich charakter za niepewny (ani zdecydowanie pozytywny, ani zdecydowanie negatywny).

Oznaczanie materiałów DF w badaniu będzie rozumiane jako umieszczenie przez twórcę informacji o nieprawdziwych treściach widocznych w materiale. To oznaczenie powinno być w miejscu widocznym dla odbiorcy, np. w tytule lub opisie publikowanego materiału albo przekazywane za pomocą cyfrowego znaku wodnego. „Cyfrowy znak wodny" to termin określający metody wdrażania jakiejś informacji do jakiegokolwiek sygnału nośnego w sposób transparentny [11]. Dobrze wdrożony cyfrowy znak wodny pozwoli na oznaczenie materiału w taki sposób, że usunięcie znaku może nie być możliwe lub spowodować zniekształcenie materiału w widoczny sposób. To sprawia, że w momencie, w którym odbiorca spotyka się z materiałem nie u jego źródła, dalej ma świadomość m.in. o prawach autorskich [11], a w przypadku DF – że jest to materiał nieprawdziwy.

3.2. DOBÓR PRÓBY BADAWCZEJ

Osobami badanymi są studenci polskich uczelni. Według raportu Feliksiaka [12] dla CBOS każda osoba w wieku od 18 do 34 lat korzysta z internetu, a około 80% z nich ogląda w internecie różnego rodzaju materiały audiowizualne. Dla porównania w kolejnych grupach wiekowych odsetek osób oglądających wideo w internecie wynosi około 50% i mniej. Jest więc bardziej prawdopodobne, że osoby młode, m.in. studiujące, zetknęły się (świadomie lub nieświadomie) lub zetkną w przyszłości z DF w formie wideo. Jest to również najstarsze pokolenie, które w pełni korzysta z internetu, dlatego to ich opinia będzie istotna w kształtowaniu się norm publikacji treści w internecie. Ponadto wśród osób w wieku 18–34 lat blisko 50% deklaruje publikowanie w internecie zrobionych przez siebie zdjęć i filmów [12]. Przy założeniu, że część tych materiałów zawiera ich wizerunki, badana grupa jest najbardziej narażona na wykorzystanie przez kogoś ich wizerunków w DF w formie wideo, do których czasami potrzeba tylko jednego zdjęcia lub filmiku [13].

Techniczne wymagania związane z uczestnictwem w badaniu to możliwość odtworzenia i obejrzenia zawartych w badaniu DF w formie wideo oraz znajomość języka angielskiego w stopniu co najmniej podstawowym, co pozwala na zrozumienie słów wypowiadanych w podanych materiałach.

3.3. TREŚĆ BADANIA

Badanie składa się z czterech materiałów wideo – każdy z nich reprezentuje jedną z kategorii analitycznych. W przypadku piątej kategorii – filmu pornograficznego – ze względów

etycznych zamiast materiału do obejrzenia umieszczone zostało pytanie projekcyjne, w którym respondent musiał wyobrazić sobie sytuację, w której natrafia na film pornograficzny ze wstawionym wizerunkiem osoby publicznej. Mimo zmiany formuły badania dla ostatniej kategorii jest ona istotna nie tylko ze względu na historię DF (ich popularność rozpoczęła się właśnie od sfałszowanych filmów pornograficznych [2]), ale również na niezwykle szkodliwy wpływ na osoby, których wizerunek wykorzystano [14].

Badanie zawiera następujące materiały:

– Projekt artystyczny: pierwsze wideo opublikowane przez konto Deeptomcruise w serwisie TikTok wykorzystujące wizerunek Toma Cruise'a. W trwającym 35 sekund materiale postać wykorzystująca wizerunek Toma Cruise'a wita się ze społecznością TikToka oraz odbija piłkę kijem golfowym. W treści materiału nie ma nic kontrowersyjnego ani też ośmieszającego Toma Cruise'a, co wpisuje się w założenia kategorii [15].
– Satyra: film opublikowany w serwisie YouTube przez użytkownika Marshall Artist o tytule *President Joe Biden's magical pistachio story (deepfake AI)*. Wykorzystuje on wizerunek prezydenta Stanów Zjednoczonych Joe Bidena i opowiada historię o magicznej pistacji, która pomogła mu odnaleźć drogę w sklepie. Wideo stylizowane jest na fragment programu informacyjnego ze stopką „Breaking News". Materiał można uznać za ośmieszający polityka ze względu na sens słów, które wypowiada, szczególnie biorąc pod uwagę środowisko, w którym się znajduje – tak jakby był w trakcie jakiegoś przemówienia [16].
– Materiał propagandowy: materiał wytworzony w ramach instalacji artystycznej *Spectre* Billa Postersa wykorzystujący wizerunek Marka Zuckerberga, który porusza kwestię wykorzystywania danych w celu przewidywania ludzkich zachowań [17]. Spośród innych materiałów pokazywanych w trakcie projektu *Spectre* ten został wybrany ze względu na rozpoznawalność osoby przedstawionej. Mimo tego, że materiał kwalifikuje się również jako projekt artystyczny, to jego treść ma zwrócić uwagę odbiorców na problemy prywatności w internecie oraz na to, na co mogą wpływać dane pozostawiane przez użytkowników internetu, dlatego spełnia kryteria materiału propagandowego.
– Fake news: DF przedstawiający fragment programu *Wolf News*, w którym prezenter mówi o częstotliwości strzelanin w Stanach Zjednoczonych, z dramatyczną muzyką w tle. Nie ma informacji, kto jest autorem tego DF, ale jest on przykładem wykorzystywania DF w kampaniach dezinformacyjnych – w tym wypadku organizacji wspierających rząd chiński przeciwko Stanom Zjednoczonym [18]. Materiał i jego treść ewidentnie mają za cel wpłynięcie na myślenie odbiorców przez rozpropagowanie fałszywych informacji.

Wymienione materiały wideo zostały wybrane również ze względu na ich dostępność oraz jakość: wyraźnie przedstawione postacie oraz dobra jakość dźwięku. Wszystkie DF są związane ze Stanami Zjednoczonymi – przedstawiane osoby są Amerykanami lub treść DF związana jest z tym krajem. Ponieważ osoby badane są studentami polskich uczelni, materiały mogą nie dotyczyć ich bezpośrednio, co pozwala zachować dystans – gdzie np. w przypadku DF przedstawiającego prezydenta Ukrainy [4] wyniki badań mogłyby być bardziej zradykalizowane, ze względu na możliwy pośredni lub bezpośredni wpływ treści oraz przedstawionych sytuacji na respondenta.

Badanie zostało zaprojektowane oraz przeprowadzone w formie CAWI (wspomaganego komputerowo wywiadu przez stronę WWW) w terminie 24–31 maja 2023 roku, za pośred-

nictwem platformy Google Forms. Kwestionariusz składał się z 9 stron – 43 pytań, w tym 3 pytań filtrujących oraz 5 pytań metrykalnych. Link do badania wraz z zaproszeniem do uczestnictwa i krótkim opisem został opublikowany na grupach na Facebooku zrzeszających studentów polskich uczelni. Zebrano 99 odpowiedzi, w tym 87 pełnych i zakwalifikowanych do dalszej analizy.

4. WYNIKI BADANIA I ICH INTERPRETACJA

4.1. POSTRZEGANIE TECHNOLOGII DEEPFAKE

Postrzeganie technologii DF zostało zbadane za pomocą czterech pytań kwestionariusza – dwóch umieszczonych na początku oraz dwóch na końcu badania. Trzeba zaznaczyć, że na odpowiedzi na pytania umieszczone na końcu scenariusza mogło wpłynąć całe badanie, w trakcie którego respondenci zostali zaznajomieni z konkretnymi materiałami DF oraz jedną hipotetyczną sytuacją użycia DF. Badani mogli odpowiadać na te dwa końcowe pytania przez pryzmat tych konkretnych przykładów, a nie DF ogółem.

Spośród badanych 58,6% zadeklarowało, że spotkało się wcześniej z DF; 17,2% nie spotkało się wcześniej z DF, a 24,1% nie miało co do tego pewności. Badani zostali również zapytani w pytaniu otwartym o to, jak rozumieją termin „deepfake". Wśród odpowiedzi znalazły się definicje terminu oraz skojarzenia z nim związane. Zostały one podzielone na dziesięć kategorii skupiających się na różnych aspektach DF i jego definicji (tab. 2). Przedstawione wyniki można podsumować następująco: 45,6%[2] respondentów postrzega DF jako technologię, a dokładniej jako proces manipulowania obrazem, wideo i/lub audio. 29,1% odpowiedzi traktowało DF jako gotowy, sfałszowany materiał lub wynik. W tym wypadku podział na DF stworzone za pomocą technologii bądź nie jest niezwykle istotny – materiały DF powstają z wykorzystaniem uczenia maszynowego i algorytmów, co odróżnia je od tzw. dumb fake'ów, które zakładają stworzenie fałszywego materiału za pomocą prostych technik montażowych [1].

Kolejna proponowana definicja DF jest bardzo ciekawa – DF jako podszywanie się pod kogoś za pomocą technologii (3,79% odpowiedzi). Istotą tej definicji jest wykorzystanie czyjegoś wizerunku, na co wskazało ponad 43% respondentów, ale nie w kontekście stworzenia jakiegokolwiek materiału (np. humorystycznego, dezinformacyjnego), a konkretnie kradzieży tożsamości. Wskazuje to na bezpośrednie skojarzenie DF z czymś złym, przestępstwem.

Ostatnia definicja, którą podało 14 respondentów (17,72%), również ma negatywne zabarwienie – fałszywa informacja, oszustwo. Odpowiedzi te mogły się wiązać z bezpośrednim przetłumaczeniem terminu „deepfake" na język polski, a tym samym mogą wynikać z braku znajomości tego terminu i niespotkania się wcześniej z DF (deklarowane przez 17,2% badanych).

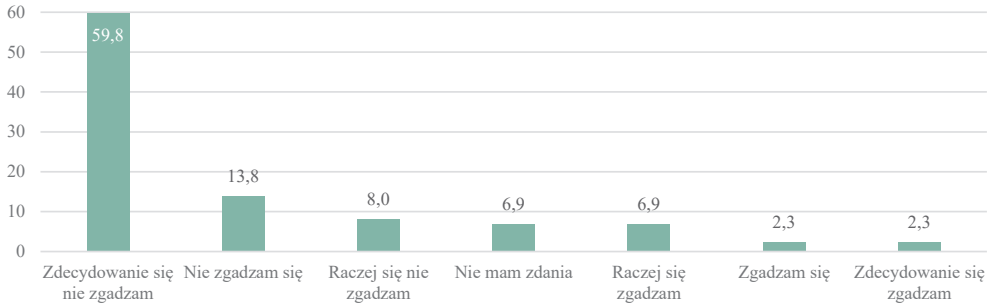
[2] Wartości procentowe w kontekście tego pytania obliczane są w stosunku do uzyskanych odpowiedzi, nie respondentów ogółem.

Tabela 2. Uogólnione definicje terminu „deepfake” na podstawie uzyskanych odpowiedzi oraz liczba respondentów odpowiadających w dany sposób

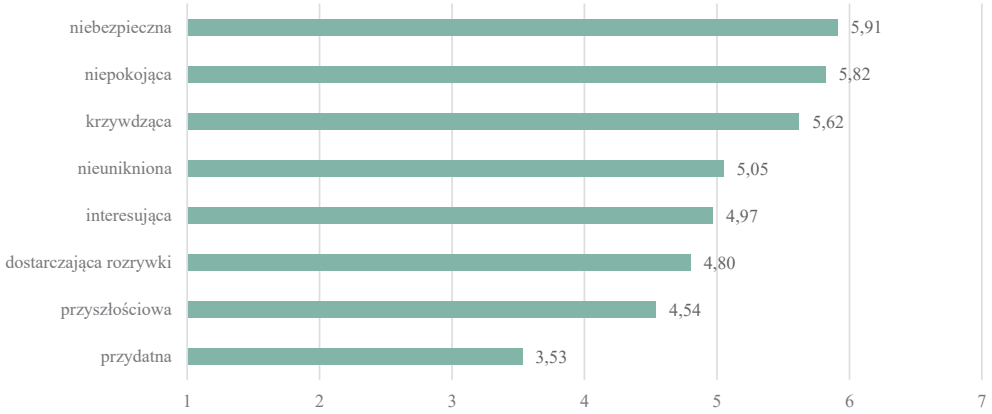
Definicja terminu „deepfake”	Liczba odpowiedzi	Procentowy wynik względem uzyskanych odpowiedzi na pytanie [%]
Technologia manipulowania obrazem, obróbka obrazu	9	11,39
Technologia manipulowania obrazem zawierającym czyjś wizerunek	9	11,39
Technologia manipulowania obrazem, wideo lub audio	9	11,39
Technologia manipulowania obrazem, wideo lub audio wykorzystującymi czyjś wizerunek i/lub głos	9	11,39
Sfałszowany materiał	4	5,06
Sfałszowany materiał przedstawiający czyjś wizerunek	7	8,86
Fałszywy materiał stworzony za pomocą technologii	3	3,79
Fałszywy materiał stworzony za pomocą technologii, przedstawiający czyjś wizerunek	9	11,39
Podszywanie się pod kogoś za pomocą technologii	3	3,79
Fałszywa informacja, oszustwo	14	17,72
Niesklasyfikowane dane	3	3,79

W pytaniu o skłonność do wykorzystania swojego wizerunku do stworzenia materiału DF (rys. 1) 59,8% respondentów zaznaczyło skrajną na podanej 7-punktowej skali odpowiedź „zdecydowanie się nie zgadzam”. Mimo tego, że w pytaniu nie zostało bezpośrednio określone, jaki to będzie charakter DF oraz jaka będzie jego treść, respondenci wyraźnie niechętnie myślą o DF ze swoim udziałem. Pokazuje to pośredni negatywny stosunek do korzystania z technologii DF.

Ten negatywny stosunek potwierdza się także w odpowiedziach na kolejne pytanie, gdzie zapytano o to, w jakim stopniu podane określenia DF zgadzają się z ich własnymi przekonaniami (rys. 2). Badani studenci, mimo tego, że uważają technologię DF za dostarczającą rozrywki oraz przyszłościową, to jednocześnie nie widzą przydatnych jej zastosowań. Uzyskany rozkład pokrywa się z widoczną tendencją w rozważaniach na temat zastosowań technologii DF – autorzy zdecydowanie częściej są w stanie wskazać zagrożenia związane z takimi materiałami niż realne korzyści lub skupiają się tylko na ich negatywnych skutkach [6, 7].



Rys. 1. Rozkład procentowy odpowiedzi na pytanie:
„Jak bardzo zgadzasz się ze stwierdzeniem: »Był(a)bym skłonny(-a)
do wykorzystania swojego wizerunku do stworzenia materiału deepfake«”?



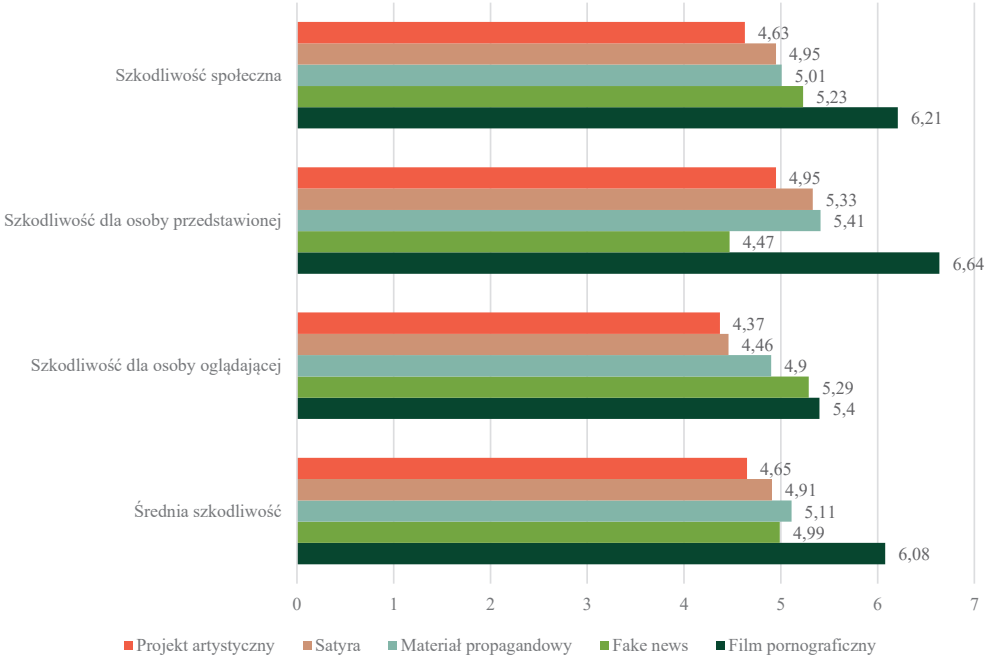
Rys. 2. Wartości średnie dla odpowiedzi na pytanie:
„Uważam, że technologia deepfake jest...”. Zastosowano skalę od 1 do 7,
gdzie 1 – „Zdecydowanie się nie zgadzam”, 7 – „Zdecydowanie się zgadzam”

Jeśli chodzi o przedstawiane materiały DF, 91,6% badanych zadeklarowało, że nie spotkało się wcześniej z żadnym z prezentowanych DF, być może dlatego, że są one związane z wydarzeniami w Ameryce, prawdopodobnie nie dotyczą ankietowanych bezpośrednio, a sami badani nie są zakładanym odbiorcą docelowym. W związku z niskim procentem osób deklarujących wcześniejsze zetknięcie się z podanym materiałem DF można uznać, że uzyskane odpowiedzi dotyczące tych materiałów (a tym samym badane opinie) w większości nie są ugruntowane i opierają się na bezpośrednim, pierwszym doświadczeniu.

Respondenci zostali poproszeni o ocenę szkodliwości każdej z kategorii materiałów DF w trzech wymiarach: szkodliwości społecznej, szkodliwości dla osoby przedstawionej oraz szkodliwości dla osoby oglądającej taki typ materiału. Na pytania badani mogli odpowiedzieć w skali od 1 do 7, gdzie 1 oznacza „Zdecydowanie szkodliwy”, a 7 – „Zupełnie nieszkodliwy”.

Porównując średnie wartości odpowiedzi, niezależnie od rozważanej sfery, można zauważyć, że respondenci postrzegali jako najbardziej szkodliwe pornograficzne materiały DF. Jeśli chodzi o szkodliwość dla bezpośredniego odbiorcy, DF pornograficzne mają jednak tylko niewielką przewagę nad fake newsami. Co ciekawe, mimo że sfałszowane filmy pornograficzne nie mają aż tak bezpośrednich skutków dla społeczeństwa w porównaniu np. z treściami dezinformującymi [6], to w przedstawianym badaniu współczynnik postrzeganej szkodliwości społecznej DF pornograficznego jest największy, i to prawie o jeden stopień większy niż w przypadku kolejnej kategorii DF. Warto jednak jeszcze raz zaznaczyć, że na wyniki mogą mieć wpływ dobrane treści – respondenci mogli oceniać konkretnie przedstawiony materiał oprócz przypadku filmu pornograficznego, gdzie postawieni zostali przed sytuacją hipotetyczną, a więc oceniali wyobrażoną przez siebie sytuację.

Jak pokazano na rysunku 3, w opinii studentów najbardziej alarmujący jest DF pornograficzny (ze średnią szkodliwością 6,08). Kolejny pod względem średniej postrzeganej szkodliwości jest materiał propagandowy (5,11), chociaż średnio szkodliwy dla osoby go oglądającej. Dalej fake news (4,99), który studenci uznają za relatywnie mało szkodliwy dla osoby przedstawionej, oraz satyra (4,91). Badani studenci za najmniej szkodliwy materiał uznają projekt artystyczny, ze średnią 4,65 – mniejszą o prawie 1,5 punktu od materiału pornograficznego.



Rys. 3. Średnie odpowiedzi na pytania dotyczące szkodliwości społecznej deepfake'ów, szkodliwości dla osoby przedstawionej oraz je oglądającej względem każdej kategorii. W pytaniach zastosowano skalę od 1 – „Zdecydowanie szkodliwy", do 7 – „Zupełnie nieszkodliwy". Na podstawie średnich wyników szkodliwości na każdej z płaszczyzn porównana została również średnia szkodliwość każdej kategorii

Ciekawy jest również rozkład średniej szkodliwości jakiegokolwiek materiału DF dla każdej z płaszczyzn. Średnia szkodliwość społeczna wynosi 5,21, dla osoby przedstawionej jest to 5,36, a dla osoby oglądającej – 4,88. Na podstawie tych wyników można stwierdzić, że DF najbardziej szkodliwe są nie dla społeczeństwa, a dla osoby, której wizerunek jest wykorzystany. Według badanych DF mają najmniej szkodliwy wpływ na osoby je oglądające. Takie podejście (największa szkodliwość dla osoby, która jest przedstawiona) widać w charakterze już istniejących praw regulujacych DF, które zakładają głównie ochronę osób, których wizerunek został wykorzystany w materiale pornograficznym, oraz osób kandydujących na stanowiska państwowe [5] (drugi rodzaj ochrony zakłada głównie ochronę demokracji oraz spokoju publicznego, ale również wizerunku). Raport, który przygotowali Huijstee i inni [5], wskazuje jednak, że najniebezpieczniejszym ze skutków DF jest ich wpływ na zaufanie społeczeństwa wobec materiałów audiowizualnych w mediach, co w kontekście tego badania wskazywałoby na większą szkodliwość dla społeczeństwa i osoby oglądającej.

Wysoka szkodliwość wobec osób, których wizerunek został przedstawiony, może mieć inne źródło – w badaniu respondentami były osoby prywatne, więc w tych wynikach potencjalnie przejawia się odczuwane przez nich zagrożenie ze strony DF. Podstawą do takiego wniosku może być również odpowiedź na wcześniej omawiane pytanie dotyczące wykorzystania swojego wizerunku do stworzenia DF, gdzie prawie 60% respondentów zdecydowanie by tego nie zrobiło. Pytanie zostało zadane na końcu badania, co znaczy, że na pewno nie ma wpływu na wcześniejsze odpowiedzi dotyczące szkodliwości, dlatego takie powiązanie może być w ogóle rozważane, chociaż nie ma na nie bezpośrednich dowodów.

4.2. OZNACZANIE FAŁSZYWOŚCI MATERIAŁU DEEPFAKE

Prezentacji każdego materiału DF towarzyszyły pytania nie tylko o postrzeganą szkodliwość, ale także o to, czy dany materiał powinien być według badanych oznaczony jako fałszywy. Najmniej kontrowersyjne rodzaje materiału DF to materiały pornograficzne (90,08% odpowiedzi za oznaczaniem ich fałszywości) oraz fake newsy (88,5% za oznaczaniem). Kolejne kategorie DF otrzymujące większość głosów za oznaczeniem fałszywości to projekty artystyczne (83,9% za oznaczaniem) oraz materiały propagandowe (82,76% za oznaczaniem), a na samym końcu, chodź z niewiele niższym wynikiem – DF o treściach satyrycznych (79,31% za oznaczaniem).

Oprócz zdecydowanych „za" lub „przeciw" oznaczaniu fałszywości materiałów DF niektórzy respondenci wybierali odpowiedź „to zależy od danego materiału". Najwięcej takich odpowiedzi uzyskał projekt artystyczny (12,64%) oraz materiał propagandowy (10,34%), a najmniej fake news (4,6%). Pokazuje to nieznaczną zależność między treścią materiału a koniecznością oznaczania go.

Rozkład odpowiedzi na pytania zadane bezpośrednio pod materiałami DF częściowo pokrywa się z rozkładem odpowiedzi uzyskanych w pytaniu umieszczonym pod koniec kwestionariusza. Widoczne jest jednak rozróżnienie w kwestii wykorzystania wizerunku między wizerunkiem osoby publicznej i osoby prywatnej. Najwięcej respondentów (86,2%) zaznaczyło, że konieczne jest oznaczenie fałszywości materiału przedstawiającego osobę prywatną, wypowiadającą się lub działającą w sposób sprzeczny z jej prawdziwymi poglądami, gdzie taki

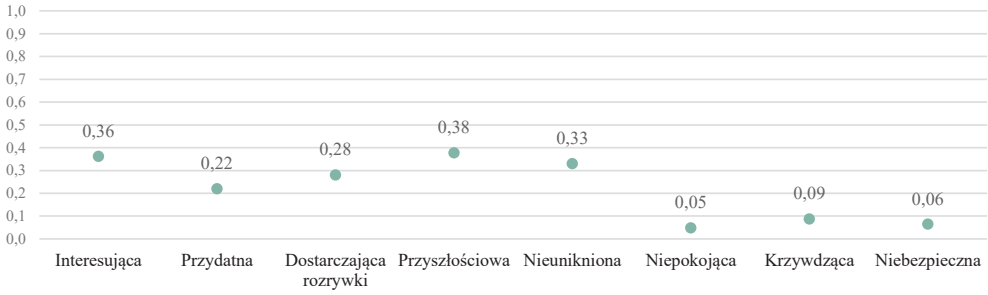
sam materiał, ale przedstawiający osobę publiczną, za obowiązkowy do oznaczenia uznaje 82,76% badanych. Inaczej w przypadku filmów pornograficznych – 85,05% określa te materiały jako obowiązkowe do oznaczenia w przypadku wykorzystania wizerunku osoby publicznej, gdzie taki obowiązek przy wykorzystywaniu wizerunku osoby prywatnej deklaruje około 3,5% respondentów mniej. 3 respondentów (3,45%) zaznaczyło zaś, że nie ma potrzeby, aby którykolwiek z podanych materiałów był oznaczany.

Ciekawe głosy znalazły się w polu otwartej odpowiedzi, w przypadku zaznaczenia opcji „Inne". Niektóre podkreślały znaczenie praw autorskich i treści wykorzystywanych do stworzenia DF, zaś jeden respondent konkretnie zaznaczył, jak powinno to oznaczenie wyglądać. Pomysł zakłada stworzenie znaku wodnego, który będzie pojawiał się na jakimkolwiek publikowanym materiale i to administracja platformy dystrybuującej dany materiał będzie określać jego ostateczną szkodliwość na podstawie zgłoszeń użytkowników. Brak oznaczenia wiązać ma się z odpowiedzialnością prawną. Trudno jednak wyegzekwować takie regulacje, jeśli autor jest nieznany lub anonimowy, a jest to częste w przypadku materiałów DF [5].

4.3. RELACJA Z TECHNOLOGIĄ A OPINIA DOTYCZĄCA TECHNOLOGII DEEPFAKE

W następnej kolejności sprawdzono, czy znajomość technologii mierzona według skali ATI-S [19] jest powiązana z opiniami dotyczącymi technologii DF. W pierwszej części omówione zostaną zależności między skalą ATI-S a opiniami na temat DF, zaś w drugiej to, czy skala ATI-S przewiduje postrzeganą szkodliwość różnych rodzajów DF (i jeśli tak, to w jakim stopniu ją przewiduje).

W celu zbadania zależności między skalą ATI-S a opiniami na temat DF sprawdzone zostały współczynniki korelacji liniowej między tymi dwoma zmiennymi. Rysunek 4 przedstawia wyniki tej analizy, zgodnie z którymi żadna z opinii na temat DF wyrażona w przymiotnikach opisujących ich możliwy charakter nie wykazuje bardzo silnej zależności liniowej ze skalą ATI-S (wszystkie współczynniki mają korelacje poniżej wartości 0,4). Wszystkie uzyskane korelacje są jednak pozytywne, wskazują na to, że im wyższa średnia skali ATI-S, tym wyższy wynik wyrażający wyższy stopień zgody z danym określeniem technologii DF.



Rys. 4. Porównanie współczynników korelacji liniowej między wynikami skali ATI-S a opiniami na temat wyszczególnionych w badaniu określeń dotyczących technologii deepfake

Najsilniejsze zależności występują między skalą ATI-S a opiniami na temat tego, czy technologia DF jest przyszłościowa (korelacja 0,38) i interesująca (korelacja 0,36). Podobnie im wyższa ogólna znajomość technologii wśród respondentów, tym mocniejsze ich przekonanie, że technologia DF jest nieunikniona (korelacja 0,33) i dostarczająca rozrywki (korelacja 0,28). Można także zauważyć, że znajomość technologii jest bardzo słabo skorelowana (prawie zupełnie nieskorelowana) z opiniami na temat tego, czy technologia DF jest niepokojąca, krzywdząca i niebezpieczna (korelacje poniżej 0,1). Wyniki te sugerują, że opinie w tych trzech kwestiach prawdopodobnie nie zależą o tego, czy ktoś lubi i umie sprawnie posługiwać się szeroko rozumianymi technologiami.

Druga część analizy skupia się na przewidywaniu postrzeganej szkodliwości DF dla różnych ich kategorii. W tym celu dla każdej kategorii DF wykorzystana została „średnia szkodliwość" – średnia wyniku postrzeganej szkodliwości społecznej dla osoby przedstawianej i oglądającej. Wartości te zostaną wykorzystane jako zmienne niezależne w serii regresji liniowych o takiej samej konstrukcji (z taką samą listą zmiennych niezależnych).

Każdy ze zbudowanych modeli regresji liniowej zawiera następujące zmienne zależne:
– średnia skali ATI-S, która określa relację respondenta z technologią, tzn. jej szeroko rozumianą znajomość (skala 1–7);
– zmienne demograficzne: wiek (wyrażony w latach), płeć (zmienna dychotomiczna, gdzie 1 – mężczyzna) i poziom studiów (zmienna dychotomiczna, gdzie 1 – studiowanie na studiach II stopnia lub podyplomowych);
– studiowanie kierunku społecznego (zmienna dychotomiczna) – studenci kierunku społecznego mogą mieć wyższe kompetencje w zakresie dostrzegania szkodliwości technologii DF zarówno dla społeczeństwa, jak i jednostki.

Wbrew oczekiwaniom żaden ze zbudowanych w ten sposób na potrzeby analizy modeli regresji liniowej nie dał wyników statystycznie istotnych (wartości F powyżej 0,05). Ze względu na słabą rozpiętość wiekową wśród respondentów (odchylenie standardowe o wartości 2,83, gdzie miarą jest wiek mierzony w latach) i osiągane wysokie wartości p w modelach dla zmiennej „wiek" zbudowana została jeszcze raz ta sama seria modeli regresji liniowych bez tej zmiennej. Wszystkie modele mają po 87 obserwacji. Tym razem otrzymano jeden model, którego wyniki są statystycznie istotne: ze zmienną zależną postrzeganej szkodliwości materiału pornograficznego ($F = 0{,}038$). Pozostałe modele nie są statystycznie istotne, nawet jeśli lista zmiennych niezależnych została zredukowana do jednej – średniej skali ATI-S. Oznacza to, że postrzeganie szkodliwości DF w czterech kategoriach (materiału artystycznego, satyry, materiału propagandowego i fake newsa) nie jest zależne od indywidualnej relacji z technologią.

Tabela 3 zawiera wyniki dla modelu dotyczącego materiału pornograficznego. Pokazuje ona, że dwie zmienne zależne – skala ATI-S oraz płeć – mają znaczenie dla przewidywania postrzeganej szkodliwości materiału pornograficznego DF (dla wartości p na poziomie niższym niż 0,05). Im niższy średni wynik w skali ATI-S, tym wyższa postrzegana szkodliwość materiału pornograficznego. Im bardziej ktoś jest obeznany z technologią w ogóle, tym mniejsza w jego oczach szkodliwość pornograficznych materiałów DF. Dodatkowo, co ciekawe, deklarowanie płci męskiej zwiększa prawdopodobieństwo wyższej postrzeganej szkodliwości

takiego materiału. Trzeba jednak zaznaczyć, że współczynnik determinacji R^2 naszego modelu wynosi tu jedynie 0,12, co oznacza, że zmienność w postrzeganej szkodliwości materiału pornograficznego wyjaśnionej przez model jest dość niska. W przyszłych badaniach warto więc poszukać innych wytłumaczeń, które mogłyby okazać się ważniejsze dla zrozumienia czynników wpływających na opinie i postrzeganą szkodliwość różnego rodzaju materiałów DF wśród studentów, ale także wśród osób w różnych kategoriach wiekowych oraz o różnym poziomie wykształcenia.

Tabela 3. Przewidywanie postrzeganej szkodliwości pornograficznego materiału deepfake – analiza regresji liniowej

	B	*SE*	*t*	Wartość *p*
(Stała)	2,609	0,460	5,671	0,000
Średnia ATI-S	**–0,182**	**0,086**	**–2,119**	**0,037**
Mężczyzna	**0,663**	**0,249**	**2,666**	**0,009**
Studia II stopnia lub podyplomowe	–0,133	0,237	–0,560	0,577
Kierunek społeczny	–0,089	0,243	–0,366	0,715
$F = 0{,}038$				
$R^2 = 0{,}12$				

5. PODSUMOWANIE I WNIOSKI

Przedstawione badanie pokazało, że studenci rozumieją DF na kilka sposobów: jako technologię (czyli proces tworzenia), materiał (czyli gotowy wynik), ogólnie pojętą fałszywą informację i próbę podszywania się pod kogoś. Stosunek do DF jest raczej negatywny, studenci uważają je za bardziej niebezpieczne i niepokojące niż przydatne i przyszłościowe. Jeśli chodzi o szkodliwość przyjętych kategorii DF (projekt artystyczny, satyra, materiał propagandowy, fake news oraz materiał pornograficzny), za najbardziej szkodliwy uznawany jest sfałszowany materiał pornograficzny, a za najmniej szkodliwy – projekt artystyczny. Poglądy dotyczące konieczności oznaczania fałszywości DF w zależności od kategorii nieznacznie się różniły, choć ogólnie opowiadano się za koniecznością ich oznaczania. Postrzeganie DF nieznacznie zależy natomiast od obycia z technologią, chociaż osoby, które deklarują większe obycie z technologią, uważają ją za bardziej przyszłościową i interesującą, a materiały pornograficzne za mniej szkodliwe.

Przeprowadzone badanie jest dobrym punktem wyjścia do dalszych poszukiwań w tym zakresie. Zaleca się jednak dobór bardziej zróżnicowanej próby badawczej nie tylko pod względem respondentów (różne grupy wiekowe), ale również treści do badania. W przypadku większego zainteresowania kwestią rozpoznawania fałszywości DF przez ich odbiorców warto przygotować próbę kontrolną zawierającą prawdziwe materiały, co stosowano już

w badaniach o takiej tematyce. W kwestii problematyki opinii warto rozbudować kategorie analityczne tak, by były bardziej szczegółowe i uwzględniały materiały o pozytywnym zabarwieniu, co sprawi, że wyniki będą bardziej zróżnicowane. Warto również wybrać więcej niż jeden materiał reprezentujący kategorię, by porównać, czy postrzegana szkodliwość nie zależy od treści przedstawionego materiału.

Niezależnie jednak od rodzaju planowanego badania i poruszanych kwestii każde z nich jest istotne dla rozwoju tej dziedziny badań. Temat DF jest stosunkowo nowy, ale technologia ta staje się coraz bardziej zaawansowana. Przeprowadzenie badania, a później publikacja wyników same w sobie będą zwracać uwagę na złożoność problematyki DF, tym samym otwierając dyskusję na temat regulacji prawnych, których ciągle jest niewiele, a które mogą być ważnym krokiem w kierunku ochrony potencjalnych ofiar technologii DF.

LITERATURA

[1] Giles K., Hartmann K., Mustaffa M., *The Role of Deepfakes in Malign Influence Campaigns*, NATO Strategic Communications Centre of Excellence, 2019.

[2] Gamage D., Ghasiya P., Bonagiri V.K., Whiting M.E., Sasahara K. (2022). *Are deepfakes concerning? Analyzing conversations of deepfakes on Reddit and exploring societal implications*, [w:] Barbosa S., Lampe C., Appert C., Shamma D.A., Drucker S., Williamson J., Yatani K. (eds.), *CHI '22: Proceedings of the 2022. CHI Conference on Human Factors in Computing Systems*, https://doi.org/10.1145/3491102.3517446.

[3] Shamook, *The Mandalorian Luke Skywalker deepfake*, YouTube, 2020, https://www.youtube.com/watch?v=wrHXA2cSpNU [dostęp: 30.06.2023].

[4] Boháček M., Farid H., *Protecting president Zelenskyy against deep fakes*, Arxiv, 2022, https://doi.org/10.48550/arXiv.2206.12043.

[5] Huijstee M., van, Boheemen P., van, Das D., Nierling L., Jahnel J., Karaboga M., Fatun M., Kool L., Gerristen J., *Tackling deepfakes in European policy* (PE 690.039), European Parliamentary Research Service, 2021, https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039.

[6] Chesney R., Citron D.K., *Deep fakes: a looming challenge for privacy, democracy, and national security*, California Law Review, vol. 107, 2018, s. 1753–1820, http://doi.org/10.2139/ssrn.3213954.

[7] Harris D., *Deepfakes: false pornography is here and the law cannot protect you*, Duke Law & Technology Review, vol. 17, 2019, s. 99–127.

[8] Dobber T., Metoui N., Trilling D., Helberger N., Vreese C., de, *Do (microtargeted) deepfakes have real effects on political attitudes?*, The International Journal of Press/Politics, vol. 26, iss. 1, 2021, s. 69–91, https://doi.org/10.1177/1940161220944364.

[9] Vaccari C., Chadwick A., *Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news*, Social Media + Society, vol. 6, iss. 1, 2020, s. 1–13, https://doi.org/10.1177/2056305120903408.

[10] Sundar S.S., Molina M.D., Cho E. (2021). *Seeing is believing: is video modality more powerful in spreading fake news via online messaging apps?*, Journal of Computer-Mediated Communication, vol. 26, iss. 6, s. 301–319, https://doi.org/10.1093/jcmc/zmab010.

[11] Arnold M., Schmucker M., Wolthusen S.D., *Techniques and applications of digital watermarking and content protection*, Artech House, Boston 2003.

[12] Feliksiak M., *Korzystanie z internetu w 2022 roku*, Komunikat z badań, nr 77/2022, CBOS, Warszawa 2022.

[13] Karnouskos S., *Artificial intelligence in digital media: the era of deepfakes*, IEEE Transactions on Technology and Society, vol. 1, iss. 3, 2020, s. 138–147, https://doi.org/10.1109/TTS.2020.3001312.

[14] Ayyub R., *I was the victim of deepfake porn plot intended to silence me*, Huffpost, 2021, https://www.huffingtonpost.co.uk/entry/deepfakeporn_uk_5bf2c126e4b0f32bd58ba316 [dostęp: 30.06.2023].

[15] Deeptomcruise, *Sports!*, TikTok, 2021, https://www.tiktok.com/@deeptomcruise/video/6932166297996233989 [dostęp: 30.06.2023].

[16] Marshall Artist, *President Joe Biden's magical pistachio story (deepfake AI)*, YouTube, 2023, https://www.youtube.com/watch?v=yVEhrIMc-ps [dostęp: 30.06.2023].

[17] Posters B. [@bill_posters_uk], *'Imagine this...' (2019)*, Instagram, 2019, https://www.instagram.com/p/ByaVigGFP2U/ [dostęp: 30.06.2023].

[18] Satariano A., Mozur P., *The people onscreen are fake. The disinformation is real*, The New York Times, 2023, https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html [dostęp: 30.06.2023].

[19] Wessel D., Attig Ch., Franke T. (2019). *ATI-S – an ultra-short scale for assessing affinity for technology interaction in user studies*, [w:] Alt F., Bulling A., Döring T. (Hrsg.), *Mensch und Computer 2019 – TAGUNGSBAND*, 2019, s. 147–154, https://doi.org/10.1145/3340764.3340766.