DOI: 10.51480/compress.2024-2.762

Katarzyna Mania
UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
katarzyna.mania@amu.edu.pl
0000-0001-8620-0238

Deepfake: nowe narzędzie dezinformacji we współczesnej fotografii

Deepfake: A new tool for disinformation in contemporary photography

ABSTRAKT

Wciąż rozwijające się technologie bez wątpienia mają niezwykły wpływ na funkcjonowanie społeczeństwa. Niosą one ze sobą zarówno ogromne możliwości, jak i wyzwania. W ciągu ostatnich kilku lat postępy w rozwijaniu zdolności sztucznej inteligencji, a zwłaszcza w dziedzinie uczenia maszynowego, doprowadziły do powstania kontrowersyjnej techniki, dzięki której człowiek zyskał nową zdolność polegającą na generowaniu przekonujących i coraz doskonalszych, lecz sztucznych treści multimedialnych, czyli deepfake'ów. Celem badania jest określenie znaczenia generowanych syntetycznych treści multimedialnych dla zjawiska rozpowszechniania dezinformacji. Przeprowadzono eksperyment, którego wyniki mają odzwierciedlać zdolność polskiego społeczeństwa do rozpoznawania fałszywych materiałów wizualnych i odróżnienia ich od prawdziwych zdjęć.

SŁOWA KLUCZOWE: NOWE MEDIA, DEEPFAKE, DEZINFORMACJA, SZTUCZNA INTELIGENCJA, FOTOGRAFIA

ABSTRACT

The constant development of technology undoubtedly has an extraordinary impact on the society. New technical solutions bring to the table both tremendous opportunities and challenges. Over the past few years, advances in artificial intelligence, especially in the field of machine learning, have led to the development of a controversial technique known as deepfake, that has given people the ability to generate convincing and increasingly sophisticated, yet artificial media content. The aim of this study is to determine the impact of such synthetic images on the phenomenon of spreading disinformation. An experiment was conducted, the results of which are intended to reflect the ability of Polish society to identify fake visual content and distinguish it from real photographs.

KEYWORDS: NEW MEDIA, DEEPFAKE, DISINFORMATION, ARTIFICIAL INTELLIGENCE, PHOTOGRAPHY

Mania, K. (2024), *Deepfake*: nowe narzędzie dezinformacji we współczesnej fotografii, ***com.press,*** 7(2), s. 26–47. DOI: 10.51480/compress.2023.7-2.762 www.compress.edu.pl

WPROWADZENIE

Współczesny świat został zdominowany przez gwałtowny rozwój technologii, który w wielu aspektach redefiniuje codzienne życie, pracę oraz interakcje zachodzące w społeczeństwie. Powołując się na koncepcję determinizmu technologicznego, można stwierdzić, że to właśnie technologia stanowi główną siłę napędową zmian społecznych, kulturowych i ekonomicznych. Zwolennicy tej teorii uważają, że rozwój technologiczny jest nieunikniony, a jego wpływ na ludzkość nieuchronny (Innis, 1999; McLuhan, 2001; Postman, 2004). Nawet mimo określonych celów stawianych sobie przez wynalazców, technologia może wykraczać poza kontrolę i intencje człowieka (Levinson, 2006). Celem artykułu jest wykazanie, jak duże zagrożenie dla społeczeństwa stwarza rozwijająca się technologia *deepfake* oraz przeprowadzenie własnych badań dotyczących zdolności rozpoznawania materiałów tworzonych za pomocą generatywnej sztucznej inteligencji. W ramach realizacji celu przeprowadzono analizę w formie badania ankietowego, w którym respondenci musieli odróżnić prawdziwe fotografie od materiałów stworzonych przy pomocy sztucznej inteligencji. Postawiono tezę, że technologia *deepfake* stanowi rosnące zagrożenie dla rzetelności informacji i autentyczności treści w mediach, a w szczególności w nowych mediach. Sformułowano również hipotezę, że ludzie bez względu na wiek mają problemy z rozpoznawaniem fałszywie wygenerowanych obrazów. Hipoteza ta została następnie poddana empirycznej weryfikacji.

SZKODLIWY WPŁYW DEZINFORMACJI

Dawniej dostarczaniem informacji w przestrzeni publicznej zajmowały się tradycyjne środki masowego przekazu, takie jak prasa, radio i telewizja. Przed publikacją proces weryfikacji treści należał m.in. do wydawców i dziennikarzy. W wyniku rewolucji cyfrowej, czyli upowszechnienia dostępu do internetu i rozwoju mediów społecznościowych, nie trzeba mieć specjalnych kompetencji ani też znać zasad netykiety, aby publikować informacje w sieci. Co więcej, we współczesnych mediach duże znaczenie zyskuje czas publikacji – pomijane są procedury weryfikacji prawdziwości czy to samej treści, czy źródeł w celu wywołania sensacji i uzyskania statusu pierwszego medium podającego daną informację do wiadomości (Babraj, 2019).

Dezinformacja jest pojęciem, które ma swoje korzenie w terminologii wywiadowczej. Po raz pierwszy użyto go w połowie XX wieku w Rosji, gdzie w 1923 roku powołano biuro dezinformacyjne. Instytucja ta powstała, aby przeprowadzać operacje mające na celu szerzenie nieprawdziwych informacji wśród społeczeństwa, co miało przynieść określone korzyści, na przykład w obszarze budowania pozytywnego wizerunku (Konieczny, 2021, s. 97). W publikacji Krajowej Rady Radiofonii i Telewizji *Fake news – dezinformacja online, próby przeciwdziałania tym zjawiskom z perspektywy instytucji międzynarodowych oraz w wybranych krajach państw UE, w tym Polski* (2020, s. 10) przeczytać można, że „dezinformacja jest działaniem celowym zmierzającym do wywołania zmian w świadomości odbiorców, zmian postaw wobec zjawisk oraz wywołania określonej reakcji społecznej, gospodarczej czy politycznej". Autorzy dokumentu wskazują również, że w związku z postępem technologicznym, zwłaszcza za sprawą rozwoju nowoczesnych mediów, w tym mediów społecznościowych, problem uległ znacznemu pogłębieniu i rozprzestrzenieniu. Zjawisko dezinformacji łączy się z występowaniem szumu informacyjnego, będącego efektem napływania licznych informacji z wielu różnych źródeł do jednostki, która próbuje je przetworzyć.

Rysunek 1. Podział zaburzeń informacyjnych.

Źródło: Derakhshan, H., Wardle, C. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Council of Europe report DGI(2017)09. s. 5.

Według grafiki zaprezentowanej w raporcie *Information disorder: Toward an interdisciplinary framework for research and policy making*, misinformacja

została zakwalifikowana jako fałszywa, malinformacja – jako szkodliwa, a dezinformacja jako fałszywa i szkodliwa. Warto jednak wspomnieć, że w komunikacie Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-społecznego i Komitetu Reginów w sprawie europejskiego planu działania na rzecz demokracji (2020, s. 18) wskazano, że w przypadku misinformacji mimo braku złych intencji ze strony nadawcy udostępniającego fałszywą informację (np. w gronie przyjaciół i rodziny), takie działanie wciąż może być szkodliwe i powodować wiele negatywnych następstw.

Dezinformacja, czyli celowe rozpowszechnianie fałszywych informacji, stała się powszechnym i niebezpiecznym zjawiskiem we współczesnym świecie. Jest ona stosowana przez różnorodne podmioty, takie jak państwa, organizacje, grupy interesów czy jednostki, by osiągnąć określone korzyści. Intencje wprowadzania dezinformacji są zróżnicowane i zależą od kontekstu oraz zamierzeń podmiotów zaangażowanych. Społeczeństwo staje przed wyzwaniem rozpoznawania i radzenia sobie z manipulacją informacją, dlatego istotne jest określenie, jakie motywacje i intencje stoją za rozpowszechnianiem fałszywych wiadomości oraz wprowadzaniem opinii publicznej w błąd. Aktywności o znamieniu dezinformacyjnym mogą być podejmowane z myślą o wywarciu wpływu na różnego typu procesy i decyzje polityczne, a także korzyści ekonomiczne. Ambicją rozpowszechniania nieprawdziwych informacji może być również spowodowanie szkody publicznej lub osobowej (Konarski, 2022).

DEEPFAKE JAKO FORMA DEZINFORMACJI W NOWYCH MEDIACH

Pojęcie *deepfake* zyskało szczególną popularność w 2017 roku za sprawą użytkownika serwisu internetowego Reddit. Internauta pod nickiem „deepfakes" publikował filmy pornograficzne z wizerunkami celebrytów i aktorów, które były sfabrykowane przy użyciu techniki głębokiego uczenia się (Dąbrowska, 2020, s. 91). Następnie udostępnił on społeczności specjalny kod – algorytm, który pozwalał na tworzenie sfałszowanych materiałów audiowizualnych również przez innych użytkowników (EPRS, 2021, s. 3), czego następstwem była fala publikacji filmów nie tylko o charakterze pornograficznym, ale również edukacyjnym i satyrycznym. Powstawały filmy z aktorami (np. Nicholasem

Cage'em) (Haysom, 2018), politykami (np. Barackiem Obamą) i innymi znanymi osobami (np. Markiem Zuckerbergiem) (Dabirian et al., 2020). Reddit podjął stanowcze działania po kilku tygodniach – zakazał stosowania takich praktyk i usunął strony z deepfake'ami. W podobnym czasie zareagowały Twitter, Discord oraz strony pornograficzne, które również ogłosiły podobne zasady (Bailey, 2018).

Termin *deepfake* jest połączeniem angielskich słów: *deep*, czyli głęboki (występującego też w złożeniu *deeplearning*, czyli głębokie uczenie się – DL), oraz *fake*, czyli fałszywy. Zazwyczaj używany jest w kontekście manipulacji istniejącymi już materiałami – obrazami, filmami i dźwiękiem, a także do nowych generowanych syntetycznie treści (Altuncu, Franqueira, Li, 2022, s. 1–2). *Deepfake* to technologia oparta na sztucznej inteligencji, która służy do tworzenia i edytowania treści wideo lub obrazów w taki sposób, aby ukazać coś, co nigdy nie miało miejsca w rzeczywistości (Young, 2019, s. 14). Ta koncepcja ma swoje początki w latach 90. XX wieku, kiedy to naukowcy rozpoczęli eksperymenty z komputerowo generowanymi obrazami (*Computer Generated Images* – CGI) w celu stworzenia realistycznie wyglądających postaci cyfrowych. Dopiero w 2010 roku technologia *deepfake* zyskała prawdziwy rozgłos za sprawą kilku czynników, które zbiegły się w czasie: dostępności dużych zbiorów danych, postępu w algorytmach uczenia maszynowego oraz powszechności potężnych zasobów obliczeniowych (Frąckiewicz, 2023). Przykładem deepfake'ów mogą być obrazy wytworzone za pomocą generatora Midjourney, działającego za pośrednictwem platformy Discord. Aplikacja pozwala na tworzenie grafik na podstawie komend tekstowych, czyli tzw. promptów, w których użytkownik może określić dokładne i szczegółowe parametry generowanych materiałów (Majchrzak, Szymkiewicz, 2023). Na rysunku 2 pokazano przykład obrazów wygenerowanych za pośrednictwem wspomnianej aplikacji. Grafiki przedstawiają papieża Franciszka, który pojawił się na premierze filmu *Barbie* – co oczywiście tak naprawdę nie miało miejsca.

Eksperci z portalu fact-checkingowego Demagog wskazują, że pojęcie *deepfake* początkowo było kojarzone jedynie z filmami, w których występowali bohaterowie ze zmodyfikowanymi twarzami. Obecnie termin ten został rozszerzony i odnosi się również do wszystkich obrazów generowanych przez sztuczną inteligencję, próbujących odwzorować prawdziwe osoby (Majchrzak, Szymkiewicz, 2023). Niektórzy specjaliści podchodzą do pojęcia o wiele szerzej i w definicji wskazują, że deepfake'iem może być nie tylko wideo, obraz, audio, ale także tekst. Ostatni z wyróżnionych elementów odnosi się do programów, które są w stanie uczyć się wzorców

językowych, generować teksty, a co za tym idzie – pisać artykuły, a nawet wiersze i piosenki. Coraz częściej wyróżnia się również *real-time* lub *live* deepfake'i, które umożliwiają modyfikację/nałożenie twarzy oraz zmianę głosu w czasie rzeczywistym, np. podczas transmisji na żywo, rozmowy telefonicznej, a nawet wideokonferencji (*A Deep Dive*, 2022).

Rysunek 2. Papież Franciszek na premierze filmu *Barbie*.

Źródło: obraz wytworzony za pomocą generatora Midjourney.

Warto zaznaczyć, że oprócz terminu *deepfake* istnieje również pojęcie *cheapfake* określające materiały tworzone za pomocą tradycyjnych technik edycji i manipulacji treści cyfrowych. Cheapfake'i nie wymagają kodowania, skomplikowanej pracy z sieciami neuronowymi ani umiejętności w post-produkcji (ID R&D, 2023, s. 3). W związku z tym ich tworzenie nie wiąże się z koniecznością wykorzystania dużych zasobów sprzętowych (Kasprzyk, 2021, s. 19). Specjaliści podkreślają, że takie formy mogą być niebezpieczne, ponieważ nie wymagają od ich twórcy nakładów finansowych czy dużego zaangażowania (ID R&D, 2023, s. 3).

Najczęściej spotykanymi przykładami deepfake'ów są fałszywe obrazy twarzy, nagrania głosowe i filmy, w których łączone są zmodyfikowane lub zsyntetyzowane elementy. Choć określenie *fake* – fałszywy kojarzy się pejoratywnie i sugeruje zmanipulowane lub sfabrykowane materiały, mające

wprowadzać odbiorcę w błąd, ta technologia ma również inne, pozytywne zastosowania. *Deepfake* może zostać wykorzystany do tworzenia sztuki, w edukacji, a także w celach rozrywkowych. Mimo że dla tej techniki zaproponowano neutralny termin – *deep synthesis*, czyli „głęboka synteza" jako alternatywę dla *deepfake*, nowa nazwa nie zdobyła popularności i nie jest powszechnie używana (Wang, 2020). Cechą charakterystyczną deepfake'ów jest ich wiralowy (wirusowy) potencjał, co oznacza, że mają one zdolność do nieprzewidywalnego i niekontrolowanego rozprzestrzeniania się w przestrzeni internetowej, czyli w mediach społecznościowych oraz na portalach informacyjnych. Stwarza to poważne zagrożenie w kontekście pogłębiania i szerzenia się zjawiska dezinformacji. Niebezpieczeństwo jest szczególnie wysokie, gdy z taką informacją zetknie się osoba skłonna (z różnych względów) do przyjęcia jej jako prawdziwej (Kulesza, Muniak, 2022).

Przykładem deepfake'a, który zdobył wielką popularność i stał się wiralem w przestrzeni internetowej, jest wytwór autorstwa Jordana Peela (Fagan, 2018). Materiał przedstawia byłego prezydenta Stanów Zjednoczonych Ameryki Baracka Obamę, który przeklina i wyzywa prezydenta Donalda Trumpa, a następnie zawraca uwagę na niebezpieczeństwo, jakie stwarza wciąż rozwijająca się zaawansowana technologia manipulacji treści.

Rysunek 3. *Deepfake* z Barackiem Obamą autorstwa Jordana Peela.

Źródło: Kadr z: BuzzFeedVideo. (2018). *You Won't Believe What Obama Says In This Video!* Youtube, https://www.youtube.com/watch?v=cQ54GDm1eL0 [data dostępu: 24.06.2024].

CHARAKTERYSTYKA PRZEPROWADZONEGO BADANIA

Dla stwierdzenia, jakie znaczenie ma deepfake'a w procesie szerzenia się dezinformacji, zdecydowano się na przeprowadzenie badania, którego celem jest określenie zdolności polskiego społeczeństwa do rozpoznawania fałszywych (wygenerowanych) materiałów i odróżnienia ich od prawdziwych fotografii. Aby ocenić słuszność postawionej hipotezy, zrealizowano analizę w formie anonimowego badania sondażowego w języku polskim, które skierowane zostało do osób powyżej 18. roku życia.

W celu klarownego przeprowadzenia analizy oraz omówienia wyników badań dokonano operacjonalizacji kilku kluczowych pojęć.

Tabela 1. Operacjonalizacja pojęć wykorzystanych w analizie.

Pojęcie	Definiowanie
Deepfake	Sfabrykowany obraz, wideo i/lub audio wygenerowany przy pomocy sztucznej inteligencji.
Technologia *deepfake*	Metody i narzędzia działające na podstawie sztucznej inteligencji, pozwalające na generowanie fałszywych treści (najczęściej obrazu, wideo i audio).
Prawdziwe zdjęcie	Fotografie wykonane przez człowieka w tradycyjny sposób, czyli za pomocą aparatów fotograficznych, smartfonów itp. Do prawdziwych zdjęć zalicza się również fotografie, które zostały poddane graficznej obróbce, czyli np. korekcie kolorów, światła, retuszowi.
Znani/nieznani	Kategorie służące do określenia osób powszechnie znanych (znani) oraz osób niepopularnych lub nieistniejących (nieznani).

Źródło: opracowanie własne.

Formularz składa się z metryczki, w której uwzględniono zmienne, takie jak: płeć, wiek, wykształcenie oraz liczba mieszkańców w miejscu zamieszkania. Zamieszczono również pakiet trzydziestu zdjęć, który składa się z dziewięciu fotografii wykonanych przez człowieka przy pomocy aparatu fotograficznego lub smartfonu oraz dwudziestu jeden wygenerowanych przeze mnie obrazów, czyli deepfake'ów. Na zdjęciach i stworzonych grafikach znajdują się ludzie popularni (kategoria: znani) oraz niepopularni lub też nieistniejący (nieznani) przedstawieni w różnych sytuacjach i okolicznościach. Celem jest zbadanie, czy jednostki potrafią zweryfikować prawdziwość zaprezentowanych materiałów. Wygenerowane zdjęcia charakteryzują się różnym stopniem trudności – jedne są „bezbłędne", czyli nie posiadają widocznych błędów, natomiast niektóre zawierają wady, o których często wspominano w mediach, np. zniekształcone ręce (Kuśmierek, 2023; Majchrzak, 2023).

Przy każdej grafice znajdowały się trzy pytania. Zadaniem respondenta było określenie prawdziwości każdego z umieszczonych w formularzu obrazów, a następnie zaznaczenie jednej z dostępnych opcji w pytaniu zamkniętym (prawda/fałsz). Za każdą prawidłową odpowiedź respondent mógł zdobyć 1 punkt. Intencjonalnie wykluczono możliwość udzielenia odpowiedzi „nie wiem" ze względu na zaistniały czynnik, jakim jest świadomość uczestniczenia w badaniu. Osoby uczestniczące w badaniu wiedziały, że niektóre z przedstawianych materiałów na pewno są spreparowane. Z tego względu respondenci podchodzili do nich zdecydowanie krytyczniej niż w przypadku codziennego przeglądania treści w mediach społecznościowych. Jeśli badany uznał zdjęcie za fałszywe, został poproszony o uzasadnienie swojej odpowiedzi. Pytano również o to, czy respondenci rozpoznają osobę na zdjęciu (pytanie zamknięte: tak/nie). Kwestia ta była bardzo istotna dla badania: po pierwsze, respondenci mogą nie znać wszystkich osób np. pełniących funkcje publiczne. Po drugie, Midjourney działa na podstawie metody głębokiego uczenia się przy pomocy generatywnych sieci współzawodniczących (*Generic Access Network* – GAN) (Schreiner, 2023). Do trenowania modelu wykorzystano miliony zbiorów danych, które są dostępne w internecie. Właściciel platformy – David Hols – tłumaczy, że nie ma sposobu, aby uzyskać sto milionów obrazów i wiedzieć, skąd pochodzą, gdyż mogą pochodzić z dowolnego miejsca. Wywołuje to wiele kontrowersji, szczególnie w środowisku fotografów, grafików i innych artystów (Growcoot, 2022).

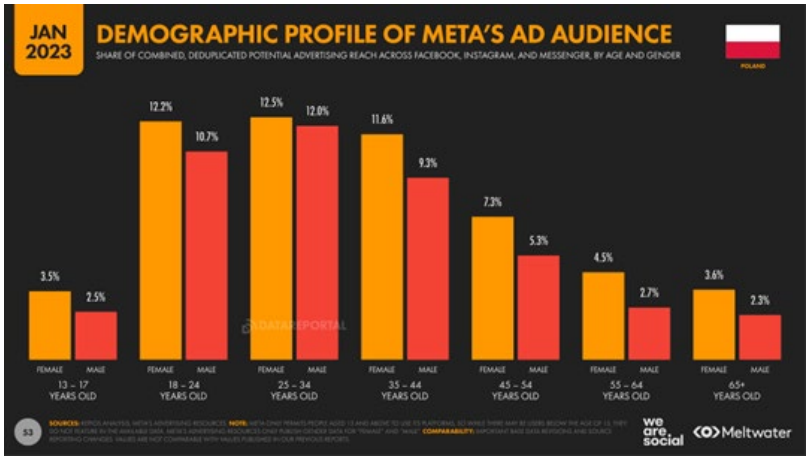
Do zrealizowania analizy niezbędne było skorzystanie z następujących narzędzi badawczych. Aby wygenerować deepfake'i, konieczne było uzyskanie dostępu do narzędzia bazującego na sztucznej inteligencji, czyli generatora obrazów, dlatego uzyskano miesięczną subskrypcję do aplikacji Midjourney. Dostęp do narzędzia jest możliwy za pomocą platformy – komunikatora Discord. Prawdziwe zdjęcia pozyskano z banku zdjęć w usłudze AdobeStock oraz z portali informacyjnych.

Badanie rozpowszechniono za pośrednictwem mediów społecznościowych, z których korzysta 27,5 miliona osób w kraju, czyli 66,3% polskiej populacji (w tym 75% osób posiadających dostęp do internetu), z czego 24,1 miliona, czyli 71% osób powyżej 18. roku życia (Digital 2023: Poland). Warto zaznaczyć, że nasilenie się zjawiska dezinformacji jest jednym z następstw rozwoju nowych mediów, w tym mediów społecznościowych (KRRiT, 2020, s. 7). Skupiono się w szczególności na Facebooku, który w Polsce jest nie tylko najchętniej wybieranym medium społecznościowym (87,5% użytkowników w przedziale wiekowym 16–64 lata), ale również najczęściej określanym

mianem ulubionego (34,5% użytkowników w przedziale wiekowym 16–64 lata) w porównaniu do innych platform (Digital 2023: Poland).

Na podstawie raportu Digital 2023: Poland przypuszcza się, że zdecydowana większość respondentów będzie kwalifikować się do przedziału wiekowego 18–34 lata ze względu na ich stosunkowo dużą aktywność w mediach społecznościowych w porównaniu do osób starszych (55–65+ lat).

Wykres 1. Podział wiekowy użytkowników mediów społecznościowych należących do firmy Meta (m.in. Facebook, Messenger, Instagram).

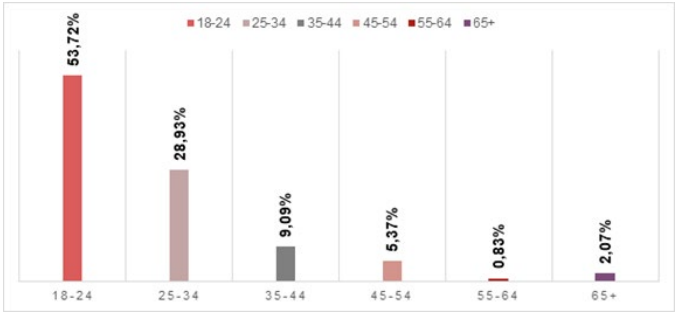


Źródło: Digital 2023: Poland (2023). https://datareportal.com/reports/digital-2023-poland, [data dostępu: 24.06.2024].

WYNIKI ANALIZY

Formularz został wypełniony przez 242 pełnoletnich anonimowych respondentów. Najliczniej reprezentowany jest segment osób w wieku 18–24 lata, stanowiąc ponad połowę (53,72%) ankietowanych. Na wykresie zauważalna jest tendencja spadkowa – osoby starsze (55–64; 65+ lat) stanowią znacznie mniejszą grupę respondentów niż młodzi dorośli, dorośli oraz osoby w średnim wieku (segment 18–54 lata).

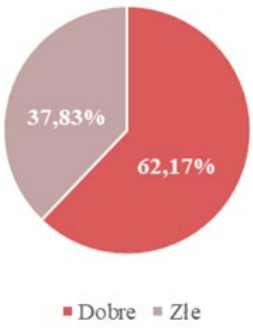
Wykres 2. Podział respondentów ze względu na wiek.



Źródło: opracowanie własne.

W ankiecie respondent mógł określić prawdziwość lub nieprawdziwość 30 zdjęć, czyli łącznie poprzez prawidłowe wskazywanie prawdziwych oraz fałszywych obrazów miał możliwość zdobycia łącznie 30 punktów (1 prawidłowa odpowiedź = 1 punkt). Średnia liczba punktów na jednego respondenta wynosi: 18,65 (62,17%).

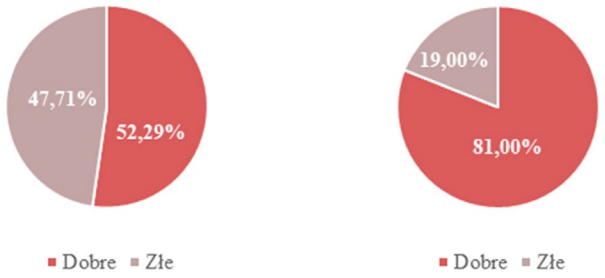
Wykres 3. Stosunek dobrych i złych odpowiedzi.



Źródło: opracowanie własne.

Odpowiedzi respondentów zostały podzielone na kategorie, które dotyczą wyłącznie wygenerowanych zdjęć deepfake'ów (fałszywe) oraz prawdziwych (prawdziwe). Średnia liczba punktów na jednego respondentach w kategorii fałszywe wynosi: 10,98 na 21 możliwych, natomiast w kategorii prawdziwe – 7,29 na 9. Zdecydowana większość odpowiedzi dotyczących prawdziwych zdjęć była prawidłowa (81%). Problemy z poprawną oceną prawdziwości przedstawionych materiałów pojawiły się przy zdjęciach wygenerowanych przy pomocy sztucznej inteligencji. Ledwo ponad połowa (52,29%) odpowiedzi była właściwa, a 47,71% wskazywało na autentyczność przedstawionych treści, które zostały sfabrykowane.

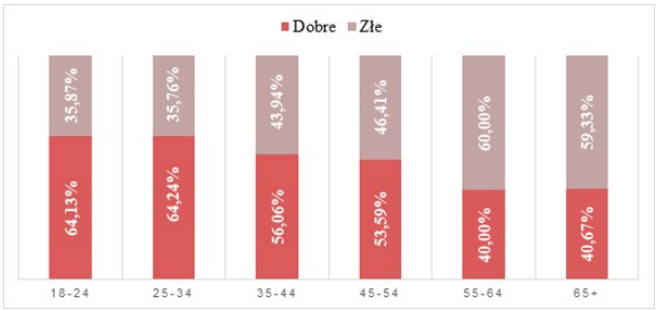
Wykres 4. Stosunek dobrych i złych odpowiedzi – kategoria fałszywe (po lewej) oraz prawdziwe (po prawej).



Źródło: opracowanie własne.

Liczbę poprawnych i złych odpowiedzi przyporządkowano i zestawiono z poszczególnymi grupami wiekowymi. Na wykresie 5 zauważyć można pewną zależność między zdolnością do rozpoznawania i poprawnego weryfikowania (w tym przypadku) treści wizualnych a wiekiem. Zaistniała tendencja spadkowa, z której wynika, że im starszy odbiorca, tym ma większe trudności z poprawną oceną autentyczności prezentowanych treści.

Wykres 5. Stosunek dobrych i złych odpowiedzi w poszczególnych grupach wiekowych.

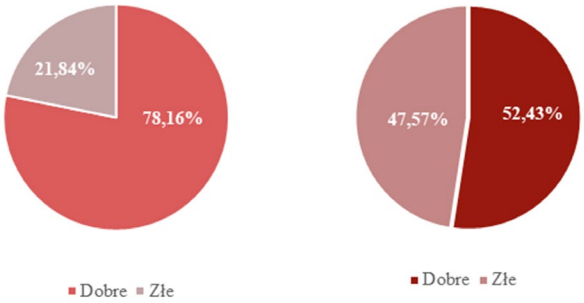


Źródło: opracowanie własne.

W ankiecie umieszczono 12 zdjęć osób, które są powszechnie znane (kategoria: znani) m.in. ze względu na pełnioną przez siebie funkcję (prezydent Stanów Zjednoczonych Ameryki Joe Biden oraz jego poprzednik Donald Trump, papież Franciszek, polski polityk i były przewodniczący Rady Europejskiej Donald Tusk, prezydent Republiki Francuskiej Emmanuel Macron oraz były premier Wielkiej Brytanii Boris Johnson). Pośród 12 zdjęć 4 z nich były prawdziwe i pochodziły z witryn internetowych, natomiast 8 było fałszywych. Reszta obrazów (18) przedstawiała osoby, które nie są znane (pobrane z banku zdjęć) lub – w przypadku sfabrykowanych fotografii – nie istniały. Dokonano zestawienia, w którym porównano stosunek odpowiedzi dobrych i złych w podziale na materiały przedstawiające wizerunki

osób powszechnie znanych oraz nieznanych lub nieistniejących (kategoria: nieznani). Z zestawienia wykluczono wybory osób, które przy odpowiedzi zaznaczyły, że nie znają postaci, która na potrzeby pracy została uznana za personę powszechnie znaną. W przypadku osób znanych padło zdecydowanie więcej prawidłowych odpowiedzi (78,16%). Z wykresie 6 można odczytać, że wyraźnie trudniejsza okazała się weryfikacja autentyczności zdjęć ludzi, którzy nie są znani jednostkom.

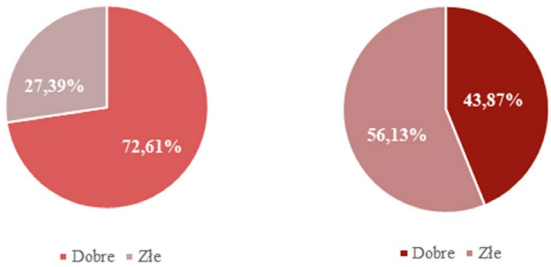
Wykres 6. Porównanie stosunków dobrych i złych odpowiedzi dotyczących wszystkich zdjęć w kategorii znani (po lewej) oraz nieznani (po prawej).



Źródło: opracowanie własne.

Kolejne wykresy (wykres 7) przedstawiają stosunek odpowiedzi dobrych i złych tylko w przypadku deepfake'ów dotyczących ludzi powszechnie znanych (znani) oraz nieznanych lub nieistniejących (nieznani). W przypadku materiałów przedstawiających wizerunek jednostek, które są kojarzone przez respondentów, większość (73,61%) była zweryfikowana prawidłowo. Zupełnie inaczej stosunek dobrych i złych odpowiedzi rozkłada się w kategorii nieznani. Większość (56,13%) odpowiedzi była nieprawidłowa, czyli wskazywała na wygenerowany za pomocą głębokiego uczenia materiał jako na autentyczny.

Wykres 7. Porównanie stosunków dobrych i złych odpowiedzi dotyczących tylko deepfake'ów w kategorii znani (po lewej) oraz nieznani (po prawej).



Źródło: opracowanie własne.

W badaniu posłużono się również metodą analizy jakościowej, której celem było przyjrzenie się spostrzeżeniom badanych jednostek. Respondenci, oprócz wskazywania prawdziwych i fałszywych obrazów, w przypadku określenia danego materiału jako „fałsz" zostali poproszeni o uzasadnienie swojej odpowiedzi.

Podczas badania jednym z aspektów, na który zwracali uwagę respondenci, był kontekst, zwłaszcza w przypadku osób powszechnie znanych. Generowane grafiki przedstawiały np. polityków w różnych sytuacjach, które nie zawsze odpowiadały pełnionym przez nich funkcjom. Przykładem jest wygenerowany *deepfake* z Joe Bidenem w różowym garniturze oraz papieżem Franciszkiem w świątyni pełnej wiernych.

Na jednym ze zdjęć (rysunek 4) prezydent Stanów Zjednoczonych jest przestawiony jako gość na premierze filmu i ma na sobie wspomniany wyżej różowy garnitur. Respondenci zauważyli, że pojawienie się Joe Bidena na premierze filmu (w tym kontekście *Barbie*), a do tego w takim ubraniu, jest bardzo mało prawdopodobne. Wielu ankietowanych wskazywało też nie tylko na brak otaczającej go ochrony, ale również obojętność osób, które znajdują się w tle zdjęcia, na obecność prezydenta. Większość ankietowanych rozpoznała deepfake'a, lecz 24,8% z nich uznało obrazek za prawdziwy.

Tabela 2. Anonimowe wypowiedzi respondentów uzasadniające nieprawdziwość obrazu z Joe Bidenem w różowym garniturze.

„Tłum patrzy w inną stronę. Nie jest typowe dla prezydenta USA by chodzic w takim stroju".
„To chyba niby Biden. Nierealna sytuacja. Prezydent w różu. Myślę, ze foto zrobiło AI".
„Wszystko wszystkim, ale Joe ma etykietę, a ta przede wszystkim stroni od tego koloru, i czerwonych dywanów. Z reszta pamiętam ze taki garnitur chyba miał Pierce Brosnan".
„Protokół dyplomatyczny nie dopuszcza takiego koloru garnituru + brak ochroniarzy".
„Joe biden nie występował w barbie".
„No błagam, nie ten odcień różu".

Źródło: wypowiedzi respondentów [zachowano oryginalną pisownię].

Rysunek 4. *Deepfake* z Joe Bidenem w różowym garniturze.

Źródło: zdjęcie wygenerowane w Midjourney.

Kolejnym przykładem jest „fotografia" papieża Franciszka w kościele (rysunek 5). Ten *deepfake* został uznany za prawdziwą fotografię przez większość ankietowanych (65,29%). Osoby, które oznaczyły zdjęcie jako fałszywe, uzasadniały swoją decyzję m.in. małym zainteresowaniem wiernych głównym bohaterem zdjęcia oraz brakiem ochrony.

Rysunek 5. *Deepfake* z papieżem Franciszkiem w kościele.

Źródło: zdjęcie wygenerowane w Midjourney.

Wśród respondentów znalazły się osoby, które zwracały uwagę na szczegóły, takie jak światło, jakość obrazu czy detale na ubraniach. Jedno ze zdjęć przedstawia kobietę podczas imprezy w klubie (rysunek 6). Ten *deepfake* został uznany za prawdziwą fotografię przez 73,14% ankietowanych. Szczególną uwagę przykuły podobne stroje uczestników imprezy, przez co jeden z respondentów określił obraz mianem „zlotu fanów Adidasa". Innymi detalami wyłapanymi przez ankietowanych była idealna cera pierwszoplanowej bohaterki zdjęcia, nieprawidłowe oświetlenie, pojawiające się refleksy oraz niewyraźne i zniekształcone części ciała postaci w tle.

Tabela 3. Anonimowe wypowiedzi respondentów uzasadniające nieprawdziwość obrazu z dziewczyną w odblaskowej kurtce.

„Mężczyzna z tyłu dziwnie trzyma kubek? Dodatkowo dużo osób ma podobne koszulki, motyw z koszulki pojawia się także na spodniach mężczyzny z tyłu".
„Oprócz elementów odblaskowych które powinny świecić, świeci również materiał na prawej ręce dziewczyny, który nie jest odblaskowy. Dół termosu/kubka chłopaka po prawej też dziwnie odbija światło".
„Ona jest zbyt idealnie oświetlona jak na imprezę w klubie".
„Rozmyte, zbyt idealne, dziwnie z układem dłoni chłopaka, który trzyma kupek, brakuje kciuka, który by trzymał".

Źródło: wypowiedzi respondentów [zachowano oryginalną pisownię].

Rysunek 6. *Deepfake* z dziewczyną w odblaskowej kurtce na imprezie.

Źródło: zdjęcie wygenerowane w Midjourney.

Warto wspomnieć, że jednym z najbardziej analizowanych elementów znajdujących się na zdjęciach były dłonie. Może to wynikać z wielu

artykułów publikowanych w portalach internetowych, które wskazywały, że jedną z największych trudności dla generatywnej sztucznej inteligencji jest prawidłowe odtworzenie rąk. Przykładem takiego deepfake'a jest grafika przedstawiająca rozmawiających Donalda Tuska oraz Emmanuela Macrona (rysunek 7). Na pierwszy rzut oka wygląda ona jak zwyczajna fotografia dwóch rozmawiających polityków. Przy wnikliwej analizie można jednak dostrzec, że prezydent Republiki Francuskiej nie ma jednego palca. Respondenci zauważyli również nieprawidłowości dotyczące nienaturalnie gładkich twarzy oraz zbyt małej ilości detali na garniturach polityków. Niektórzy z odpowiadających dostrzeli także sygnety na dłoni Tuska i wyrazili wątpliwość w to, czy prezes Rady Ministrów nosi taką biżuterię.

Rysunek 7. *Deepfake* przedstawiający Donalda Tuska i Emmanuela Macrona z brakującym palcem.

Źródło: zdjęcie wygenerowane w Midjourney.

PODSUMOWANIE

Z przeprowadzonych badań wynika, że *deepfake* stanowi realne i wciąż rosnące zagrożenie dla społeczeństwa. Fascynująca i budząca spore zainteresowanie generatywna technologia stwarza wiele niebezpieczeństw, które wymagają natychmiastowego działania w celu wypracowania środków zapobiegawczych. Ta technologia z łatwością może być wykorzystywana przez jednostki, organizacje i państwa do osiągania korzyści z rozprzestrzeniania dezinformacji, a także do wyrządzania szkód społecznych i politycznych. Należy mieć na uwadze, że *deepfake* może znaleźć również nieszkodliwe, pozytywne zastosowania, m.in. w produkcjach filmowych i branżach kreatywnych. Co więcej, może ona ułatwić przygotowanie dubbingu w różnych językach przy zachowaniu odpowiedniej mimiki twarzy lub generowania realistycznie wyglądających obrazów bez konieczności przeprowadzania czasochłonnych i kosztownych sesji zdjęciowych. Ponadto technologia *deepfake* może wspierać edukację, umożliwiając np. tworzenie interaktywnych materiałów z udziałem wirtualnych postaci historycznych.

Przeprowadzone badanie potwierdziło hipotezę, że ludzie, niezależnie od wieku, mają problemy z odróżnianiem deepfake'ów od prawdziwych zdjęć. Niepokojącym wynikiem jest odsetek odpowiedzi uznających umieszczone w formularzu deepfake'i za zgodne z prawdą, ponieważ, jest to aż 47,71%, czyli prawie połowa. Należy zaznaczyć, że w ankiecie znalazły się również wygenerowane przez Midjourney obrazy, które zawierały widoczne błędy, takie jak: nieprawidłowa liczba palców, przesadnie wygładzone twarze, nienaturalne sylwetki ludzi znajdujących się na drugim planie lub też całkowicie nierealny kontekst, a mimo to część respondentów nie dostrzegła tych niedoskonałości i uznawała materiały za zgodne z rzeczywistością. Największe trudności zaobserwować można u osób starszych, u których wystąpiły problemy ze zweryfikowaniem autentyczności przedstawionych obrazów, co oznacza, że są oni bardziej podatni na dezinformację rozpowszechnianą w formie deepfake'ów. Okazało się, że najwięcej deepfake'ów, które uznawane były za prawdziwe zdjęcia, ukazywało wizerunki osób nieistniejących. W przypadku jednostek powszechnie znanych respondenci częściej demaskowali wygenerowane przez sztuczną inteligencję obrazy poprzez doszukiwanie się nieprawidłowości w kontekście sytuacyjnym widocznym na obrazie bądź też szczegółów w wyglądzie osoby będącej bohaterem przedstawionej grafiki (np. kształt głowy).

Omawiając przeprowadzoną analizę, trzeba mieć na uwadze, że jednym z istotnych czynników, który niewątpliwie miał wpływ na finalne wyniki,

jest świadomość bycia badanym. Respondenci wypełniali formularz, przypuszczając, że na pewno natkną się na sfabrykowane obrazy, dlatego często doszukiwali się nieprawidłowości nawet w prawdziwych zdjęciach. Można śmiało stwierdzić, że taka czujność powinna występować także podczas codziennego przeglądania treści w przestrzeni internetowej, ponieważ to właśnie tam jesteśmy szczególnie narażeni na zetknięcie się z dezinformacją. Lepiej mieć wątpliwości co do prawdziwej informacji i podchodzić do niej z dystansem niż zostać zmanipulowanym. Dlatego warto przywołać słowa Stanisława Lema: „lepiej mieć zero informacji, niż być zdezinformowanym, bo oznacza to informację ujemną, mniejszą od zera" (Lem, 2012). Należy również pamiętać, że technologia jest narzędziem w rękach człowieka. O tym, czy możliwość szybkiego generowania materiałów multimedialnych stanowi zagrożenie, czy też nowe szanse i możliwości dla społeczeństwa, decydują ludzkie motywy i intencje.

BIBLIOGRAFIA

Altuncu, E., Franqueira, V. N., Li, S. (2022). *Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review*. arXiv preprint, arXiv:2208.10913.

Babraj, R. (2019). *Czym jest fact-checking? – zarys inicjatyw na świecie i w Polsce*. NASK. CyberPOLICY. Pobrano z: https://cyberpolicy.nask.pl/czym-jest-fact-checking-zarys-inicjatyw-na-swiecie-i-w-polsce/#_ftn9 (dostęp: 10.06.2024).

Bailey, J. (2018). *The deepest fake: how new tech will test our belief in what we see*. The Sydney Morning Herald. Pobrano z: https://www.smh.com.au/technology/the-deepest-fake-how-new-tech-will-test-our-belief-in-what-we-see-20180423-p4zb4w.html (dostęp: 10.06.2024).

BuzzFeedVideo. (2018). *You Won't Believe What Obama Says In This Video!* Youtube. Pobrano z: https://www.youtube.com/watch?v=cQ54GDm1eL0 (dostęp: 10.06.2024).

Communication From The Commission to The European Parliament, The Council, The European Economic and Social Committee and The Committee of The Regions on the European democracy action plan. (2020). European Commission. COM(2020) 790 final. Pobrano z: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0790&from=EN (dostęp: 10.06.2024).

Dabirian, A., Kietzmann, J., Whittaker, L., Kietzmann, T. C. (2020). *All around Me Are Synthetic Faces': The Mad World of AI-Generated Media*. 99.

Dąbrowska, I. (2020). Deepfake, nowy wymiar internetowej manipulacji. *Zarządzanie Mediami*, *8*(2), 89–101.

Digital 2023: Poland. (2023). Pobrano z: https://datareportal.com/reports/digital-2023-poland (dostęp: 10.06.2024).

EPRS. (2021). *Tackling deepfakes in European policy.* European Parliamentary Research Service.

Fagan, K. (2018). *A viral video that appeared to show Obama calling Trump a ‚dips---' shows a disturbing new trend called ‚deepfakes'*. Business Insider. Pobrano z: https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4?IR=T (dostęp: 10.06.2024).

Frąckiewicz, M. (2023). *The Evolution of Deepfake: Tracing the History and Development of AI-generated Content*. TS2 SPACE.

Growcoot, M. (2022). *Midjourney Founder Admits to Using a 'Hundred Million' Images Without Consent*. PetaPixel. Pobrano z: https://petapixel.com/2022/12/21/midjourny-founder-admits-to-using-a-hundred-million-images-without-consent/ (dostęp: 10.06.2024).

Haysom, S. (2018). ‚*People Are Using Face-Swapping Tech to Add Nicolas Cage to Random Movies and What Is 2018*,' Mashable,
Za: EPRS. (2021). *Tackling deepfakes in European policy.* European Parliamentary Research Service Scientific Foresight Unit (STOA). PE 690.039.

ID R&D (2023). *Deep Dive into Deepfakes. Mitigating the growing threat to biometric security posed by fake digital imagery and injection attacks*. Pobrano z: https://www.idrnd.ai/wp-content/uploads/2023/03/IDR_D_Deepfakes_Whitepaper.pdf (dostęp: 10.06.2024).

Innis, H. A. (1999). *The Bias of Communication, Toronto–Buffalo.*

Konarski, X. (2022). *Dezinformacja online – jak ją rozumieć i jakie są środki prawne jej zwalczania w Polsce i UE*. Pobrano z: https://www.traple.pl/dezinformacja-online-jak-ja-rozumiec-i-jakie-sa-srodki-prawne-jej-zwalczania-w-polsce-i-ue/ (dostęp: 10.06.2024).

Konieczny, M. K. (2021). Wokół pojęcia dezinformacji w aspekcie prawno-kryminalistycznym. *De Securitate et Defensione. O Bezpieczeństwie i Obronności 7*(2), 95–108.

KRRiT (2020). *Fake news – dezinformacja online. Próby przeciwdziałania tym zjawiskom z perspektywy instytucji międzynarodowych oraz wybranych państw UE, w tym Polski.*

Kulesza, W., Muniak, P. (2022). *Jesteśmy przekonani, że potrafimy zdemaskować deepfake. Badania pokazują coś odwrotnego.* Newsweek. Pobrano z: https://www.newsweek.pl/zdrowie-i-nauka/nauka/deepfakes-ludzie-nie-potrafia-zdemaskowac-deepfakeow-jak-temu-zaradzic/mn02dxn (dostęp: 10.06.2024).

Kuśmierek, M. (2023). *Zabierze ci pracę, ale za to wygeneruje osiem krzywych palców. Dlaczego SI jest beznadziejna w tworzeniu dłoni?* Spider'sWeb. Pobrano z: https://spidersweb.pl/2023/04/sztuczna-inteligencja-generowanie-dloni.html (dostęp: 10.06.2024).

Lem, S. (2012). *Opowieści o pilocie Pirxie*. Kraków: Wydawnictwo Literackie.

Levinson, P. (2006). *Miękkie Ostrze, czyli historia i przyszłość rewolucji informacyjnej*. Warszawa: MUZA.

Levinson, P. (2010). *Nowe nowe media*. Kraków: Wydawnictwo WAM.

Majchrzak, A., Szymkiewicz, A. (2023). *Złowieszczy Obama i stylowy papież – ewolucja deepfake'ów.* Demagog. Pobrano z: https://demagog.org.pl/analizy_i_raporty/zlowieszczy-obama-i-stylowy-papiez-ewolucja-deepfakeow/ (dostęp: 10.06.2024).

McLuhan, M. (2001). Media i zmiany kulturowe. W: McLuhan E., Zignore F. (red.), *McLuhan. Wybór tekstów* (s. 124–135). Poznań: Zysk i S-ka.

Postman, N. (2004). *Triumf Techniki nad Kulturą*. Warszawa: MUZA.

Schreiner, M. (2023). *GigaGAN: An old AI architecture shows off some new tricks*. Decoder. Pobrano z: https://the-decoder.com/gigagan-an-old-ai-architecture-shows-off-some-new-tricks/ (dostęp: 10.06.2024).

Turbani, R. (2023). *Deepfakes: como conteúdo falso alimentado por IA pode distorcer a percepção da realidade*. Época Negócios. Pobrano z: https://epocanegocios.globo.com/tecnologia/noticia/2023/04/deepfakes-como-conteudo-falso-alimentado-por-ia-pode-distorcer-a-percepcao-da-realidade.ghtml (dostęp: 10.06.2024).

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., … & Ling, Z. H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, 101–114.

Young, N., (2019). DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media, New York , s. 14. Za: Dąbrowska, I. (2020). Deepfake, nowy wymiar internetowej manipulacji. *Zarządzanie Mediami*, *8*(2), 89–101.