WOJSKOWA AKADEMIA TECHNICZNA

im. Jarosława Dąbrowskiego

ROZPRAWA DOKTORSKA

Bartosz BIDERMAN

Zagrożenie dezinformacją multimedialną z wykorzystaniem technologii deepfake z perspektywy bezpieczeństwa narodowego

Promotor:
prof. dr hab. inż. Tadeusz SZCZUREK

Promotor pomocniczy:
dr Jakub ADAMKIEWICZ

Warszawa 2024

Streszczenie

Dysertacja przedstawia wyniki procesu badawczego ukierunkowanego na rozwiązanie problemu zawartego w pytaniu: „Czy technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu?”. Praca składa się ze wstępu, jednego rozdziału metodologicznego oraz pięciu rozdziałów merytorycznych. Zwieńczona jest zakończeniem oraz bibliografią, stanowiącą główne źródło wiedzy teoretycznej. Załączony został również spis ilustracji, tabel, wykresów oraz arkusz czterech załączników zawierających metryczkę z badania, jej dodatkowe analizy statystyczne, kwestionariusz badawczy oraz skróty statystyczne użyte w pracy.

Rozdział pierwszy zawiera metodyczne podstawy opracowania – uzasadnienie wyboru tematu, cel i przedmiot badań oraz opis założonych problemów badawczych i hipotez. Przedstawione zostały zastosowane metody badawcze, takie jak analiza danych zastanych oraz eksperyment badawczy, a następnie opisano założenia i ograniczenia badawcze.

Rozdział drugi zawiera omówienie teoretycznego wpływu technologii deepfake na jednostki i ich postrzeganie rzeczywistości. Przedstawiona została definicja kluczowych pojęć i wyniki dotychczasowych badań. Przeprowadzono analizę teoretyczną, w jaki sposób fałszywe nagrania mogą zakłamywać percepcję i wpływać bezpośrednio na bezpieczeństwo personalne, a pośrednio na bezpieczeństwo strukturalne, jako składowe funkcjonowania państwa i społeczeństwa.

W rozdziale trzecim opisane zostały techniczne aspekty tworzenia nagrań deepfake. Szczegółowo przedstawiono proces tworzenia nagrań, potrzebne zasoby oraz niezbędną wiedzę techniczną. W rozdziale przedstawiono wyniki badań wyjaśniające, jak powstają nagrania deepfake i jakie narzędzia są do tego potrzebne.

Rozdział czwarty zawiera omówienie wyników eksperymentu pod względem zdolności internautów do rozpoznawania zmanipulowanych treści. Analizuje, jak jednostki reagują na fałszywe materiały multimedialne i jakie czynniki wpływają na percepcję manipulacji.

W rozdziale piątym przeanalizowane zostały wyniki badania eksperymentalnego pod względem wpływu nagrań deepfake na bezpieczeństwo personalne. Opisano, w jaki sposób filmy deepfake wpływają na decyzje osób w kontekście oszustw finansowych.

Istotą rozdziału szóstego jest przedstawienie rekomendacji dotyczących przeciwdziałania zagrożeniom związanym z technologią deepfake, ze szczególnym naciskiem na aspekty praktyczne. Zaproponowane zostały narzędzia oraz strategie przeciwdziałania działaniom z wykorzystaniem technologii deepfake w obszarach takich jak dezinformacja, oszustwa personalne, zastraszanie lub próby kompromitacji danej osoby czy organizacji.

We wnioskach do poszczególnych rozdziałów i w zakończeniu podsumowano główne ustalenia pracy badawczej oraz przedstawiono wyniki weryfikacji postawionych hipotez. Wskazano na rosnącą potrzebę opracowania mechanizmów obronnych, zarówno w wymiarze technicznym, jak i edukacyjnym. Podkreślono również konieczność prowadzenia dalszych interdyscyplinarnych badań nad technologią deepfake oraz ich wpływem na społeczeństwo, zwłaszcza w kontekście manipulacji i dezinformacji. Wskazano, iż analiza wpływu fałszywych nagrań na zaufanie społeczne, procesy decyzyjne, rynki finansowe i stabilność polityczną może dostarczyć cennych informacji na temat skali i natury zagrożenia.

Słowa kluczowe: zagrożenia bezpieczeństwa narodowego, bezpieczeństwo personalne, bezpieczeństwo strukturalne, nagrania deepfake, manipulacja obrazem, dezinformacja.

Abstract

The dissertation presents the results of a research process aimed at addressing the question: „Does deepfake technology enable the generation of realistic recordings that can influence the decisions of those who view them, thereby posing a threat to national security?”. The work consists of an introduction, one methodological chapter, and five substantive chapters. It is concluded with a summary and a bibliography, which serves as the primary source of theoretical knowledge. Additionally, a list of illustrations, tables, graphs, and a sheet of four appendices are included, containing the survey's metadata, additional statistical analyses, the research questionnaire, and statistical abbreviations used in the study.

The first chapter presents the methodological foundations of the study, including the justification for the topic selection, research objectives and subject, and a description of the assumed research problems and hypotheses. The applied research methods, such as secondary data analysis and experimental research, are outlined, along with the research assumptions and limitations.

The second chapter discusses the theoretical impact of deepfake technology on individuals and their perception of reality. It introduces key definitions and reviews existing research findings. A theoretical analysis is conducted on how fake recordings can distort perception and directly affect personal security and, indirectly, structural security, as components of state and societal functioning.

The third chapter describes the technical aspects of creating deepfake recordings. It provides a detailed explanation of the process, required resources and necessary technical knowledge. The chapter presents research findings explaining how deepfake recordings are made and what tools are needed.

The fourth chapter discusses the results of the experiment concerning the ability of Internet users to recognize manipulated content. It analyzes how individuals react to fake multimedia materials and what factors, for example psychological and personality traits, influence the perception of manipulation.

The fifth chapter analyzes the experimental findings regarding the impact of deepfake recordings on personal security. It describes how deepfake videos affect individuals' decisions in the context of financial fraud.

The essence of the sixth chapter is to provide recommendations for countering the threats posed by deepfake technology, with a particular emphasis on practical aspects. Tools and strategies for counteracting activities using deepfake technology in areas such

as disinformation, personal fraud, intimidation, or attempts to compromise an individual or organization are proposed.

In the conclusions of each chapter and the final section, the main findings of the research are summarized, and the results of hypothesis verification are presented. The growing need for the development of defensive mechanisms, both technical and educational, is highlighted. The necessity for further interdisciplinary research on deepfake technology and its societal impact, especially in the context of manipulation and disinformation, is emphasized. The analysis of the influence of fake recordings on public trust, decision-making processes, financial markets, and political stability may provide valuable insights into the scale and nature of the threat.

Keywords: national security threats, personal security, structural security, deepfake recordings, image manipulation, disinformation.

Spis treści

Wprowadzenie

Temat omawiany w pracy jest niezwykle istotny i aktualny, dotykający kluczowych wartości, takich jak poczucie bezpieczeństwa i wpływ nagrań na zaufanie społeczne. Rozwój technologii deepfake oraz innych metod manipulowania obrazem i dźwiękiem, to dziedziny nowe, intensywnie rozbudowywane. Z jednej strony pozwala to badaczom na aktywne partycypowanie w rozwoju technik manipulacji, z drugiej zaś wiąże się z wieloma zagrożeniami, takimi jak ciągła dezaktualizacja tematu, trudności w dostępie do literatury, ograniczona liczba ekspertów i brak zweryfikowanych metodologicznie badań dotykających tej tematyki.

Stosunkowo niewielka liczba badań behawioralnych sprawdzających wpływ fałszywych nagrań na zachowanie i decyzje jednostek potwierdza aktualność tematu, szczególnie w kontekście bezpieczeństwa narodowego. Nieliczne analizy poświęcone zostały pogłębianiu wiedzy na temat różnic w identyfikacji i podatności na fałszywe nagrania w zależności od cech osobowych jednostki oraz jej doświadczeń życiowych. Ponadto brakuje danych o wpływie materiałów deepfake na decyzje podejmowane przez osoby pełniące funkcje publiczne, które padają ofiarą manipulacji, co może wpłynąć pośrednio na bezpieczeństwo narodowe.

Kolejnym elementem przemawiającym za wyborem tematu pracy jest jego bezdyskusyjna przynależność do dyscypliny nauk o bezpieczeństwie. Termin deepfake występuje w wielu polskojęzycznych opracowaniach z tej dziedziny, jednak samo pojęcie ma wiele niespójnych definicji. Deepfake jako technologia niesie ze sobą nie tylko wysoki potencjał, lecz także wymagające zidentyfikowania zagrożenia dla bezpieczeństwa państwa i jego obywateli. W trakcie prowadzonych badań ustalono te zagrożenia, opisano je i zaproponowano schemat rozwiązań mający na celu zmniejszenie podatności społeczeństwa na fałszywe nagrania.

Udało się to dzięki zastosowaniu badań opartych na rozwiązaniach, typowych dla nauk o bezpieczeństwie lub szerzej nauk społecznych. Przede wszystkim wykorzystano metody empiryczne, takie jak eksperyment badawczy i analiza danych zastanych. Ponadto korzystano z różnych metod teoretycznych, takich jak analiza, porównanie, abstrahowanie i wnioskowanie.

XXI wiek stanowi czas gwałtownych przemian technologicznych oraz rozwoju usług z nich korzystających. Wiele z najnowocześniejszych rozwiązań przynosi

pozytywne zmiany, lecz praktycznie każda nowa technologia generuje również negatywne efekty. W tym kontekście deepfake z pewnością wykorzystywany bywa jako element dezinformacji i manipulacji społeczeństwem, rzutując bezpośrednio lub pośrednio na stan bezpieczeństwa narodowego.

Deepfake swoją genezę zawdzięcza głębokiemu uczeniu, czyli inaczej deeplearning-owi. Stanowiące element tej kategorii uczenie maszynowe (*machine learning*) pozwala na analizy obszernych baz danych (*big data*), swobodne przetwarzanie ich oraz dokonywanie dotychczas niewyobrażalnie skomplikowanych obliczeń. Algorytmy głębokiego uczenia maszynowego są wykorzystywane obecnie między innymi jako ochrona przed spamem, ale także, jako narzędzie personalizowania tego spamu czy reklam wyświetlanych danemu użytkownikowi. Głębokie uczenie maszynowe to rodzaj zaawansowanej sztucznej inteligencji, który koncentruje się na budowie i treningu sztucznych sieci neuronowych. Opiera się na idei, że komputery mogą „uczyć się" na podstawie przykładów, bez konieczności wprowadzania do nich szczegółowych instrukcji. Pozwala to algorytmom komputerowym na analizowanie i rozumienie złożonych danych, w tym obrazów, dźwięków, tekstu i innych rodzajów informacji. Głębokie sieci neuronowe składają się z wielu warstw neuronów, które przetwarzają dane w hierarchiczny sposób, co pozwala im na identyfikowanie wzorców i cech abstrakcyjnych w danych.

W głębokim uczeniu maszynowym istnieją dwa główne rodzaje algorytmów: uczenie nadzorowane i uczenie bez nadzoru. Algorytmy uczenia nadzorowanego są używane wówczas, gdy mamy dostęp do przykładów z prawidłowymi odpowiedziami, a celem jest nauczenie systemu rozpoznawania tych odpowiedzi na nowych danych. Algorytm uczenia bez nadzoru używany jest wówczas, gdy nie ma dostępu do prawidłowych odpowiedzi i naszym celem jest znalezienie pewnych wzorców lub grup w danych. To drugie podejście jest wykorzystywane przy tworzeniu nagrań deepfake – twórca ma do dyspozycji pewne elementy i informacje wsadowe, natomiast efekt końcowy nie jest mu znany. Sieć przeciwstawna mówi twórcy, jak bardzo jego dzieło różni się od oryginału, którego nigdy nie widział.

Głębokie uczenie jest stosowane w wielu różnych dziedzinach, takich jak rozpoznawanie mowy, rekomendacje w sklepach internetowych, diagnoza chorób czy autonomiczne samochody. Algorytmy uczenia maszynowego są coraz częściej używane w przemyśle, ponieważ pozwalają na szybsze i dokładniejsze podejmowanie decyzji,

eliminując czynnik ludzki. Wykorzystuje je się również w analizach danych, programowaniu, jak i coraz powszechniej przy obróbce zdjęć czy nagrań wideo. Dzięki uczeniu maszynowemu możliwe jest automatyzowanie wielu procesów, co pozwala na oszczędność czasu i pieniędzy.

Mankamentem głębokiego uczenia maszynowego jest często konieczność posiadania dużych ilości danych początkowych oraz potężnych komputerów, sprawnych do przetwarzania i analizy ogromnej ilości informacji. Wymagana jest również często duża wiedza z zakresu informatyki, statystyki i matematyki, aby móc poprawnie zaprojektować i przeprowadzić proces nauki. Algorytmy uczenia maszynowego są podatne na błędy, jeśli zostaną wyuczone na złych lub niepełnych danych. Ponadto głębokie uczenie może być kontrowersyjne ze względu na możliwe implikacje etyczne i społeczne, takie jak brak przejrzystości w działaniu algorytmów czy ryzyko wykluczenia ludzi z procesów decyzyjnych.

Ogrom danych, jakie jest w stanie przetworzyć sztuczna inteligencja, pozwala nie tylko na analizy tekstu czy liczb, lecz także rozpoznawanie i tworzenie obrazów. Dzięki szybkiemu przyrostowi mocy obliczeniowej takowe przetworzenia odbywać się mogą obecnie w czasie rzeczywistym, nie wymagając oczekiwania na efekt końcowy. Technologia ta pozwala na przykład w trakcie wideo rozmowy nałożyć na wizerunek osoby rozmawiającej odpowiednie filtry, upiększając lub urozmaicając obraz. Głębokie uczenie wykorzystuje się również np. W automatycznym rozpoznawaniu chorób płuc poprzez komputerowe rozpoznawanie zdjęcia RTG. W 2020 roku utworzone zostały projekty[1] pozwalające automatycznie zidentyfikować przy pomocy takiego zdjęcia wirus COVID-19 oraz stan jego zaawansowania[2].

Po powyższych przykładach zdawać się może, iż technologia głębokiego uczenia maszynowego wpływa na rozwój społeczeństwa wyłącznie w pozytywny sposób. Jednak wraz z popularyzacją technologii część osób postanowiła wykorzystać ją do swoich własnych celów. Pod koniec 2017 roku w Internecie pojawiły się nagrania prezentujące znane aktorki kinowe w trakcie czynności seksualnych. Nagrania były sfałszowane, a do ich utworzenia wykorzystano głębokie uczenie maszynowe. Od pseudonimu

[1] L. Brunese, F. Martinelli, F. Mercaldo, A. Santone, „Machine learning for coronavirus covid-19 detection from chest x-rays", Procedia Computer Science, Volume 176, 2020.
[2] M. Aras Ismael, Abdulkadir Şengür, „Deep learning approaches for COVID-19 detection based on chest X-ray images", Expert Systems with Applications, Volume 164, 2021.

pierwszego autora, publikującego nagrania na portalu Reddit[3] powstało określenie technologii zmieniającej obraz rzeczywistości – *deepfake*.

Wizja świata, w którym fałszywy obraz przykrywa rzeczywistość od dawna wykorzystywana jest w literaturze. Za przykład niech posłuży motyw występujący w prozie Stanisława Lema, którego jeden z bohaterów – Ijon Tichy – spacerując ulicami Nowego Jorku, wraz z każdą fiolką zwalczającą działania Federalnego Zarządu Psyprecji zastawał otoczenie zgoła odmienne od tego pierwotnego[4]. Zakłamanie rzeczywistości na przestrzeni lat postąpiło już tak daleko, iż mało kto zdawał sobie sprawę z zachodzących procesów. Były to jedynie pojedyncze jednostki, kierujące „lepszą" rzeczywistością, dostarczaną przeciętnemu obywatelowi. Reszta społeczeństwa funkcjonowała w rzeczywistości z nałożonym na niego „kamuflażem", zakrywającym brzydotę i upadek rzeczywistego świata. Obraz świata kreowany przez Stanisława Lema w "Kongresie Futurologicznym" obrazuje czytelnikowi ponurą wizję fantasy, w której człowiek nie ma wpływu na to, w jakim stopniu jego rzeczywistość jest zakłamywana. Główni bohaterzy funkcjonują w utopijnej rzeczywistości, nie mając nawet świadomości bycia okłamywanym.

Technologia *deepfake* działa podobnie. Umożliwia okłamywanie osób oglądających filmy, co do prawdziwości postaci w niej występujących. W 2019 roku, chińskiej streamerce „Her Royal Highness Qiao Biluo" w trakcie transmisji na żywo, na skutek awarii technicznej, zostały wyłączone nakładane na jej twarz filtry[5]. Na skutek awarii setki tysięcy obserwujących ją osób dowiedziało się, iż zamiast nastoletniej, „uroczej, młodej bogini", po drugiej stronie ekranu transmisję prowadzi 58-letnia kobieta.

Wraz z rozwojem technologii coraz trudniej jest zidentyfikować prawdziwe elementy obrazu i określić, co jest prawdą, a co fałszem. Technologia deepfake pozwala tworzyć cyfrową rzeczywistość zgoła odmienną od tej realnie uchwyconej na nagraniu czy fotografii. Oprócz twórcy danego dzieła oraz grupki ekspertów wyposażonych we „fiolki odkłamujące" – odpowiednie aplikacje dekonspirujące – coraz mniej osób jest w stanie odróżnić fałszywy przekaz od prawdziwego.

[3] Reddit – serwis internetowy przedstawiający linki do różnorodnych informacji, które ukazały się w Internecie. https://www.reddit.com/ [dostęp: 01.01.2023].
[4] S. Lem, „Kongres futurologiczny", 1983.
[5] Artykuł opisujący przypadek chińskiej vlogerki „Your Highness Qiao Biluo", https://www.bbc.com/news/blogs-trending-49151042 [dostęp: 01.01.2023].

Stoimy zatem u progu świata, w którym być może nie będziemy umieć odróżnić prawdy od fałszu. Informacje krążące w cyberprzestrzeni będą wymagały weryfikacji za pomocą specjalistycznych narzędzi, niedostępnych dla każdego. Zresztą, jak już zostało wspomniane, wyzwanie to jest aktualne nawet teraz. Warto zatem zastanowić się, jak bardzo przeciętny człowiek podatny jest na manipulację obrazem i dźwiękiem oraz w jakim stopniu skutecznie rozpoznaje oszustwo kryjące się w zafałszowanym przekazie. Zagadnienie to stało się przewodnim motywem niniejszej rozprawy. Zaś przedmiotem prowadzonych badań jest dezinformacja multimedialna powstała z wykorzystaniem technologii *deepfake* i jej wpływ na bezpieczeństwo narodowe.

Zidentyfikowane zostały dwa cele badawcze – cel główny poznawczy oraz cel praktyczny (utylitarny). Celem głównym jest identyfikacja zagrożeń dla bezpieczeństwa narodowego, wynikających z rozwoju technologii deepfake, zaś celem utylitarnym wskazanie kierunków działania w celu ochrony przed dezinformacją. Dzięki wyznaczeniu celu utylitarnego praca będzie miała również wymiar praktyczny. Wyniki eksperymentu bezpośrednio przełożą się na wyznaczenie elementów pozwalających na identyfikację fałszu. Pozwoli to na zaproponowanie utworzenia wystandaryzowanego pakietu edukacyjnego obywatelskich umiejętności i kompetencji w zakresie rozpoznawania użycia technologii deepfake.

W pracy przedstawiony został główny problem badawczy oraz sześć problemów cząstkowych (szczegółowych). Odpowiedź na każdy z nich znajduje się w przypisanym im rozdziale pracy. Wszystkie razem odpowiadają zaś na główne pytanie, które brzmi: Czy technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu? Pytania szczegółowe zaprezentowano poniżej, wraz z ich podziałem zastosowanym na odpowiadające im rozdziały, natomiast ich pełne rozwinięcie znajduje się w pierwszym rozdziale, podrozdział 3.3.

Próbą odpowiedzi na pytanie główne oraz pytania szczegółowe jest hipoteza główna oraz zestaw hipotez roboczych, zwanych również hipotezami cząstkowymi lub pomocniczymi. Hipoteza główna zakłada, iż technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu. Hipotezy robocze zostały zaprezentowane poniżej, wraz z kategoryzacją ich na poszczególne rozdziały.

Szczegółowy opis hipotez, przedstawiony jest w rozdziale pierwszym, podrozdział 3.4 oraz we wstępie do odpowiadających im rozdziałów.

Każde pytanie badawcze znajduje swoją domniemaną odpowiedź w odpowiadającej mu hipotezie. Każdemu z problemów szczegółowych oraz próbie jego rozwiązania poświęcony został osobny rozdział. Praca zawiera więc w sobie pięć rozdziałów merytorycznych (problemowych) oraz jeden rozdział – metodologiczny. Schemat pracy z podziałem na rozdziały wraz z zaprezentowanym przedmiotem badania, celem cząstkowym, problemem i hipotezą przedstawiony został poniżej.

Rozdział pierwszy, nazwany „konceptualizacja badań nad manipulacją obrazem" jest standardowym rozdziałem metodologicznym. Przedstawia sytuację problemową – uzasadnia konieczność podjęcia badań oraz prezentuje ich cel i przedmiot. W dalszej części rozdziału rozpisane zostały poszczególne problemy badawcze oraz założone hipotezy im odpowiadające. Przedstawiono zastosowane przy ich weryfikacji teoretyczne oraz empiryczne metody badawcze. Opisane zostały klasyczne założenia dwóch głównych zastosowanych metod badawczych – analizy dokumentów oraz eksperymentu badawczego. Następnie przedstawiono założone ograniczenia badawcze oraz szczegółowo opisano obszar, organizację i przebieg badań. Pod koniec uwzględniono zastosowane w badaniu empirycznym moderatory.

W rozdziale drugim, o tytule „Wpływ zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe – ustalenia terminologiczne" omówiono główną część teoretyczną pracy, w kolejnych punktach przedstawiono próby definiowania kluczowych określeń użytych w tytule, problemie głównym oraz w hipotezach. Następnie dokonano ustalenia i prezentacji wyników dotychczasowych badań z obszaru tematyki poruszanej w pracy. Opisano teoretyczny wpływ nagrań oraz obrazów, wytworzonych przy pomocy technologii deepfake na ocenę sytuacyjną danej jednostki oraz jej działanie. W ciągu badań ustalono jak i w jakim stopniu nieprawdziwe medium jest w stanie wpłynąć na jednostkę, tym samym tworząc u niej fałszywe przekonania, zakłamanie rzeczywistości. Celem badań, których wyniki przedstawiono w rozdziale drugim, było poszukiwanie odpowiedzi na pytanie, jaki jest status ontologiczny wiedzy na temat wpływu zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe. Weryfikowana hipoteza brzmi następująco – zmanipulowane materiały audiowizualne mogą mieć istotny wpływ na bezpieczeństwo narodowe, a obecny stan wiedzy na ten temat jest niepełny i nadal rozwijający się.

Przedmiotem badań w trzecim rozdziale, nazwanym „Techniczne aspekty nagrań deepfake oraz rozwój technologii" są metody tworzenia nagrań deepfake. Celem niniejszego rozdziału było zbadanie tego, jak przebiega tworzenie nagrań – jak bardzo jest czasochłonne, w jakim stopniu wymagana jest wiedza techniczna lub jak dużych zasobów systemowych potrzeba do utworzenia przykładowego nagrania. Opisano również szczegółowo sam proces tworzenia materiałów na potrzeby eksperymentu prezentowanego w niniejszej pracy. Pytanie badawcze trzeciego rozdziału brzmi: w jaki sposób powstają nagrania deepfake? Weryfikowana hipoteza zakłada, iż nagrania deepfake powstają poprzez wykorzystanie oprogramowania do zmiany obrazu lub dźwięku w oryginalnym nagraniu, tak, aby wyglądało to, jakby ktoś inny mówił lub wyglądał inaczej niż w rzeczywistości.

Przedmiotem badań w czwartym rozdziale, o tytule „Percepcja zmanipulowanych przekazów wizualnych w świetle procesów psychologicznych i społecznych", jest zdolność polskich internautów do weryfikacji rzetelności prezentowanych im materiałów audiowizualnych. Celem badania jest ustalenie tego, czy internauci rozróżniają materiały multimedialne prawdziwe od fałszywych. Próbą odpowiedzi na niniejsze pytanie badawcze jest następująca hipoteza: internauci częściowo rozróżniają materiały multimedialne prawdziwe od fałszywych, wytworzonych przy wykorzystaniu technologii deepfake, ale mogą mieć trudności z odróżnieniem prawdziwych materiałów od fałszywych, zwłaszcza jeśli są one dobrze wykonane. W większości przekonani są oni o ich niepodważalności i prawdzie, ponieważ nie dostrzegają minimalnie widocznych uchybień w obrazie, w zależności od przekonań są w stanie uwierzyć w przekazywane im treści. Przyjmuje się, że osoby w różny sposób odbierają podobne przeżycia czy doświadczenia. Przedmiotem badania jest więc zbadanie wpływu nagrań deepfake na wybory badanych. Celem badania jest ustalenie, czy zmanipulowane wideo może wpłynąć w istotny sposób na wybory ludzi.

Pytanie badawcze postawione w rozdziale piątym, o tytule „Wpływ nagrań deepfake na bezpieczeństwo personalne w świetle przeprowadzonego eksperymentu" brzmi: jaki wpływ ma percepcja zmanipulowanych nagrań wideo na wybory i opinie internautów, w świetle bezpieczeństwa narodowego? Weryfikowaną hipotezą zaś twierdzenie: percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów, a jej oddziaływanie na bezpieczeństwo narodowe jest

zależne od stopnia wiarygodności tych nagrań oraz od kontekstu, w jakim są one prezentowane.

Ostatni rozdział, o tytule „Przeciwdziałanie dezinformacji wyzwaniem dla bezpieczeństwa strukturalnego" dotyczy bezpośrednio poszczególnych cech osób i ich wpływu na podatność manipulacjom oraz możliwości obrony przed nią. Składa się on z dwóch części. W pierwszym fragmencie rozdziału celem badania jest ustalenie, czy istnieją poszczególne cechy, które w znacznym stopniu przyczynić się mogą do wzrostu lub spadku podatności na dezinformację deepfake. Pytanie szczegółowe prezentowane w niniejszym rozdziale brzmi: Jaki wpływ mają zastosowane moderatory na odbiór dezinformacji deepfake, a przez to na bezpieczeństwo narodowe? Zaproponowana hipoteza brzmi: osoby z podwyższonymi poszczególnymi wskaźnikami społecznymi są bardziej podatne na uleganie manipulacjom dezinformacji deepfake, co przyczynia się do zwiększenia ryzyka zagrożenia bezpieczeństwa narodowego.

W drugiej części szóstego rozdziału poruszono praktyczny aspekt badań. Przedmiotem analiz jest model przeciwdziałania nieprawdziwym mediom audiowizualnym oraz minimalizacji podatności społeczeństwa na rozwój dezinformacji prowadzonej z użyciem technologii deepfake. Celem jest prezentacja propozycji mających na celu zwiększenie świadomości społecznej oraz uodpornienie jej na prawdopodobny nagły przyrost nieprawdziwych lub wprowadzających odbiorcę w błąd nagrań.

Pytanie badawcze mające pomóc w ustaleniu punktu wyjściowego oraz osiągnięcie zaprezentowanego celu pośredniego, brzmi następująco: jakie należy podjąć działania w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake? Ustalenie owych części składowych pozwoli na analizę stanu każdego z nich oraz propozycję rozwiązań mających na celu ich poprawę. Weryfikowana hipoteza zakłada, iż aby ochronić się przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, należy podjąć działania zapobiegające rozprzestrzenianiu się takich materiałów, edukować społeczeństwo w zakresie rozpoznawania dezinformacji audiowizualnej oraz zapewnić odpowiednie narzędzia do weryfikacji prawdziwości takich materiałów.

Na potrzeby pracy zastosowano liczne ograniczniki – czasowe (listopad 2017 – czerwiec 2022), przedmiotowe (pominięcie manipulacji audiowizualnej ze względu na kontekst użycia materiału), warsztatowe (pominięto aspekty prawne, badaniu poddani

zostali wyłącznie studenci polskich uniwersytetów), technologiczne (niemożność użycia technologii śledzenia wzroku) oraz źródłowe (pominięto bogatą literaturę chińską). Ich szczegółowe omówienie znajduje się w rozdziale metodologicznym w punkcie 3.2.3.

W pracy zastosowano liczne metody badawcze, zarówno empiryczne jak i teoretyczne. Kluczowe dla poniższej pracy jest przeprowadzone badanie w schemacie eksperymentalnym wewnątrz osobowym. Zebrane przy jego pomocy dane stały się podstawą do użycia metod teoretycznych – analizy, syntezy, wnioskowania. Jednak podstawą umożliwiającą przeprowadzenie eksperymentu był pogłębiony przegląd literatury, zwany również inaczej analizą danych zastanych. Zastosowanie tej metody pozwoliło określić schematy pracy oraz zbudować hipotezy będące odpowiedzą na szczegółowe pytania badawcze.

Dobór metod badawczych nie był przypadkowy, a za jego zastosowaniem przemawia wiele elementów. Metoda analizy dokumentów[6] stanowi swoisty wstęp do dalszych badań, a jej zastosowanie utwierdziło autora w przekonaniu, iż sformułowane przez niego pytania nie znajdują w niej swych odpowiedzi. Po głębszej analizie ustalone zostało, iż ze względu na brak odpowiednich danych, koniecznie jest wywołanie określonych procesów, wprowadzenie do nich nowego czynnika, a następnie obserwowanie zachodzących zmian. Jest to definicja eksperymentu, jaką zaproponował Władysław Zaczyński[7]. W dalszej części rozważań nad proponowanymi w pracy metodami dookreślone zostało, iż ze względu na szczególną tematykę badań oraz ich wysoką wrażliwość społeczną, przeprowadzony eksperyment będzie miał charakter laboratoryjny. Jak zauważa Leszek Korzeniowski, założenie to odmienne jest od eksperymentu naturalnego. Z jednej strony uczestnik badania może nie czuć się w pełni swobodnie, jak dzieje się w przypadku eksperymentu naturalnego, jednak niewątpliwą zaletą laboratoryjnych warunków jest fakt istnienia praktycznie identycznych warunków dla każdego badanego oraz możliwość zastosowania dodatkowych elementów obserwacji (na przykład *eyetracking*)[8].

W badaniach wykorzystano w większości angielskojęzyczne artykuły naukowe. Z racji aktualności tematu, jego świeżości oraz dzięki szybkiemu rozwojowi, w nauce

[6] M. Cieślarczyk, Z. Chojnacki, „Techniki i narzędzia badawcze stosowane w pracach magisterskich i doktorskich", w: „Metody, techniki i narzędzia badawcze oraz elementy statystyki stosowane w pracach magisterskich i doktorskich", M. Cieślarczyk (red.), Warszawa 2006, s. 60.
[7] W. Zaczyński, „Praca badawcza nauczyciela", Warszawa 1995, s. 87-89.
[8] L. F. Korzeniowski, „Podstawy nauk o bezpieczeństwie", Wydanie 2, 2017, s. 41 – 52.

brakuje zwartych publikacji poświęconych tej problematyce. Ponadto w tym temacie przeważa zachodni trend, stawiający na krótkie, umieszczone w otwartym dostępie (*open access*) artykuły. Elementem wpływającym na owy trend może być fakt, iż zdecydowana większość dostępnej literatury to techniczne opisy działania technologii *deepfake* oraz informatyczne prezentacje jej właściwości i możliwości rozwoju. Nie sposób wymienić tutaj dominujących w tym obszarze autorów publikacji, nie mówiąc już o naukowych autorytetach. Wynika to z faktu, iż w przytaczanych publikacjach często znajduje się więcej niż trzech autorów, a ich wkład w powstanie pracy jest trudny do określenia. Ponadto zauważono wzmacnianie się trendu do powstawania międzyuczelnianych konsorcjów badawczych z nastawieniem na ich interdyscyplinarność. Coraz częściej zdarza się, że w jednym artykule widnieją obok siebie nazwiska informatyka, psychologa i analityka danych. W pracy wykorzystano również literaturę polskojęzyczną, zwłaszcza opisując metodologię badań i przy definiowaniu najważniejszych pojęć. Wynika to z faktu, iż niniejsza dysertacja, jak i samo badanie powstało w języku polskim. W analizie literatury przedmiotu powołano się na liczne autorytety z zakresu nauk o bezpieczeństwie, doceniając ich wybitny wkład w rozwój tej dyscypliny naukowej oraz pośrednie przyczynienie się do powstania tej dysertacji.

Rozdział 1. Konceptualizacja badań nad manipulacją obrazem

Dezinformacja stanowi jeden z głównych oręży XXI wieku[9]. Państwa demokratyczne, półdemokratyczne oraz państwa pozornie demokratyczne zobligowane są do przestrzegania szeregu traktatów i umów międzynarodowych, co w znacznej mierze ogranicza możliwość prowadzenia tradycyjnych operacji wojskowych. Ponadto decydentów zdaje się przerażać wizja wojny konwencjonalnej, prowadzonej bez żadnych ograniczeń. Ogrom zniszczeń dotknąłby nie tylko wojskowych, lecz również cywilów, a także negatywnie wpłynął na środowisko naturalne[10]. Z tych powodów na znaczeniu zyskały działania pozamilitarne, psychologiczne, takie jak na przykład dezinformacja społeczeństw czy tworzenie i rozpowszechnianie „*fake news-ów*". Jak zaznacza Kamil Mroczka, narastanie zjawiska globalnej dezinformacji jest fenomenem niezaprzeczalnym, wyraźnym, generującym realne, już ziszczone polityczne niebezpieczeństwa[11]. Państwowe, międzynarodowe organizacje i osoby prywatne stają się świadome wykorzystywania fałszywych i wprowadzających w błąd informacji, zwłaszcza w erze społeczeństwa informacyjnego. Jak zauważa K. Mroczka, według badania Eurobarometru z 2018 r. ponad 85% Europejczyków uważało, że fałszywe wiadomości były problemem w ich kraju. Nieco mniej, bo 83%, stwierdziło, że fake newsy stanowią zagrożenie dla demokracji[12]. Celem rozpowszechniania fałszywych informacji może być wywołanie określonych skutków politycznych (np. pożądany wynik wyborów, zwycięstwo jednej z opcji lub niszczenie odmiennej), ekonomicznych (np. upadek dużego banku, spekulacje giełdowe, wahania kursów) lub społecznych (np. pogłębiająca się polaryzacja nastrojów społecznych w odniesieniu do określonych kwestii). Cyberterroryści również wykorzystują Internet i manipulację zamieszczanymi tam treściami, jako kanał komunikacji do mobilizacji swoich zwolenników. Zjawisko to jest dotkliwe do tego stopnia, że w wymiarze globalnym mówi się o „rozstroju informacyjnym" (*information disorder*)[13]. Udokumentowano, iż w latach 2008 – 2017 partie polityczne i rządy wydały na kampanie dezinformacyjne ponad pół miliarda

9 V. Volkoff, Psychosocjotechnika, dezinformacja – oręż wojny, Komorów 1999, s. 40-42.
10 T. Szczurek, M. Górnikiewicz, „Social Media Wars – The R-evolution Has Just Begun", Warszawa 2018, s. 30 – 35.
11 K. Mroczka, „Fake news as a new category of threat to the system of economic security of the state in the era of epidemic crisis", Przegląd Bezpieczeństwa Wewnętrznego nr. 26 (14), 2022, s. 338-341.
12 European Union, Eurobarometer 464: „Fake News and Disinformation Online", 2018, str. 4.
13 C. Wardle, H. Derakshan, „Information Disorder Toward an interdisciplinary framework for research and policymaking", Council of Europe report DGI, 2019.

dolarów amerykańskich, a manipulacja informacją w mediach społecznościowych, zmieniła się w wysoce sprofesjonalizowany i znakomicie finansowany sektor[14]. Powszechnie wykorzystuje się w walce politycznej boty, trolle, fake-i czy inne narzędzia służące polaryzacji społeczeństwa.

Kwestią poruszaną w większości rozważań naukowców, jest to, jak dezinformacja wpływa na obywateli oraz na jedność narodu i jego zdolności obronne. Równie często zadawane pytanie dotyczy tego, w jaki sposób przeciwdziałać dezinformacji – czy to systemowo, działaniami edukacyjnymi czy na poziomie legislacji. Niemniej istotnym problemem jest rozstrzygnięcie na poziomie mikro – czy każda jednostka jest podatna na działanie dezinformacji w jednakowy sposób oraz co przyczynia się do wzrostu lub obniżenia podatności? Odpowiedź na to pytanie pozwoliłaby wyznaczyć elementy wpływające na obniżenie percepcji fałszywego przekazu, a także podniesienie kontrdezinformacyjnych możliwości. Poprzez zwiększenie samoświadomości jednostki potencjalnie zmniejszyć można podatność społeczeństwa na dezinformację.

Świat nauki od lat stara się udzielić mniej lub bardziej precyzyjnych odpowiedzi na te pytania. Debata uczonych jeszcze w pierwszych latach XXI wieku zdawała się jednak koncentrować wyłącznie wokół tradycyjnych from przekazu oraz przekazywanej tam dezinformacji – radia, telewizji czy gazet i ich propagandowej treści. Jednak dopiero Internet – wschodzące medium pozyskiwania informacji – w pełni pozwoliło wykorzystać techniki manipulacji oraz dotrzeć z nimi zarówno do pojedynczego odbiorcy jak i całych grup społecznych.

Obszarem nowym i wciąż wymagającym przeprowadzenia badań, jest dezinformacja zmanipulowanymi materiałami audiowizualnymi, tak zwanymi filmami *deepfake*, powstałymi przy wykorzystaniu technologii uczenia maszynowego. Zagadnienia wymienione powyżej zostały zawężone przez autora do materiałów tworzonych przy pomocy tej technologii. W niniejszym rozdziale omówiona została metodologia przeprowadzonych badań oraz przyjęty przez autora warsztat pracy.

14 L. M. Neudert, „Future Elections May Be Swayed by Intelligent, Weaponized Chatbots", MIT Technology Review, 2018.

1.1 Założenia metodologiczne badania wpływu zmanipulowanych materiałów audiowizualnych

W niniejszym podrozdziale opisany został zaprojektowany proces badań nad zmanipulowanymi materiałami audiowizualnymi. Dokonano uzasadnienia zastosowania wybranych technik badawczych oraz opisano założenia metodologiczne, metodyczne oraz techniczno-organizacyjne realizacji procesu badawczego. Określono motywy wyboru technik pomiaru i ich charakterystyki, wielkość próby, jej dobór oraz selekcję przypadków. Podrozdział zawiera również refleksję dotyczącą błędów pomiaru oraz wyzwań związanych z badaniem. Podrozdział zawiera ponadto – zgodnie z wymogami i regułami sztuki – analizę sformułowanej tezy, pytania badawczego oraz będących odpowiedzią – hipotez badawczych.

Zdaniem Jerzego Brzezińskiego o dojrzałości dyscypliny empirycznej świadczy to, w jakim stopniu formułowane w jej obrębie hipotezy sprawdzane są na drodze eksperymentalnej[15]. Nauki o bezpieczeństwie, swoją dojrzałością ustępują nie tylko fizyce czy matematyce, lecz także swoim „starszym kolegom" z dziedziny nauk społecznych: socjologii czy psychologii, współcześnie nadal nazywanej dyscypliną „rozwijającą się". Stąd na potrzeby niniejszej publikacji, główna metoda badawcza – eksperyment – zaczerpnięta została od starszej dyscypliny – psychologii społecznej – gdzie metoda ta została dogłębnie opisana i jest powszechnie wykorzystywana.

Jeszcze kilkanaście lat temu jako dominujące metody badawcze w naukach o społeczeństwie wymieniano wywiady ankietowe oraz pogłębione wywiady indywidualne. Na skutek ewolucji dziedziny oraz popularyzacji innych metod empirycznych, od początku drugiej dekady XXI wieku obserwuje się powolne odejście od tych metod badawczych na rzecz innych metod, takich jak eksperyment[16]. Stanisław Sulowski zauważa, że odejście od tradycyjnych metod badawczych uniemożliwia ustalanie związków przyczynowych między zmiennymi, co stawia przed badaczem kolejne wyzwania[17].

15 J. Brzeziński, „Metodologia badań psychologicznych", Warszawa 2004, s. 282.
16 J. Czaputowicz, „Czy interdyscyplinarność jest właściwym kierunkiem rozwoju stosunków międzynarodowych w Polsce?", [w:] A. Gałganek, E. Haliżak, M. Pietraś (red.), Wieloi interdyscyplinarność nauki o stosunkach międzynarodowych, Polskie Towarzystwo Stosunków Międzynarodowych, Wydawnictwo Rambler, Warszawa 2012, s. 240–242.
17 S. Sulowski, „O rozwoju badań i postulacie interdyscyplinarności w naukach o bezpieczeństwie", [w:] Sulowski S. (red.), „Tożsamość nauk o bezpieczeństwie", Toruń 2015, s. 21 – 23.

Na potrzeby weryfikacji hipotez pracy i odpowiedzi na postawione pytania, przeprowadzony został eksperyment – badanie w schemacie eksperymentalnym wewnątrz osobowym[18], określany również eksperymentem laboratoryjnym, w anglojęzycznej terminologii *controlled experiment*[19]. Jako wstęp do badań przeprowadzona została analiza danych zastanych, inaczej zwaną również badaniami zza biurka[20], a po angielsku nazywaną *secondary data analysis*[21], *retrospective data analysis*[22], *desk research*[23], *nonreactive research*[24] lub *unobtrusive research*[25].

1.1.1 Sytuacja problemowa – uzasadnienie podjęcia badań

Temat, jaki poruszono w pracy, jest niezwykle ważny i aktualny. Rozwój technologii deepfake odbywa się równolegle do procesu powstawania pracy, co z jednej strony stwarza możliwość nieustannego pogłębiania wiedzy w dziedzinie, z drugiej zaś ciągnie za sobą liczne zagrożenia, takie jak szybka dezaktualizacja tematu, trudność z dostępem do literatury, mała liczba ekspertów oraz brak zweryfikowanych metodologicznie czy prospektywnych badań nad tematem.

Co ważne, aktualność tematu potwierdza brak badań behawioralnych, badających wpływ fałszywych nagrań na zachowanie i decyzje jednostek. W nielicznej literaturze nie sposób jest również znaleźć badania sprawdzające różnice w identyfikacji oraz podatności na fałszywe nagrania w zależności od cech osobowych jednostki oraz jej orientacji na skrajne poglądy. Brakuje również badań dotyczących wpływu zmanipulowanych technologią *deepfake* nagrań na bezpieczeństwo narodowe.

---

[18] J. M. Brzeziński, „Metodologia badań psychologicznych", Warszawa 2019, s. 241 – 288.
[19] S. L. Chow, „Experimentation in psychology—rationale, concepts, and issues", [w:] „Methods in Psychological Research – Encyclopedia of Life Support Systems" (EOLSS), Eolss Publishers, Oxford, UK, 2002.
[20] Z. Bednarowska, „Desk research — wykorzystanie potencjału danych zastanych w prowadzeniu badań marketingowych i społecznych", [w:] „Marketing i Rynek" 7/2015, Kraków, s. 18 – 26.
[21] Melissa P. Johnston, „Secondary Data Analysis: A Method of which the Time Has Come", [w:] „Qualitative and Quantitative Methods in Libraries (QQML)" 2014, s. 620 – 625.
[22] R. D. Neal, D. A. Lawlor, V. Allgar, „Missed appointments in general practice: retrospective data analysis from four practices", [w:] „British Journal of General Practice", 2001, s. 829 – 832.
[23] Z. Bednarowska, „Desk research – wykorzystanie potencjału danych zastanych w prowadzeniu badań marketingowych i społecznych", Uniwersytet Jagielloński w Krakowie "Marketing i rynek", 7/2015, s. 18 – 26.
[24] E. J. Webb, D. T. Campbell, R. D. Schwartz, L. Sechrest, „Unobtrusive measures: Nonreactive research in the social sciences", Chicago, IL: Rand McNally, 1966.
[25] T. Araujo, P. Neijens, „Unobtrusive Measures for Media Research", [w:] J. Van den Bulck (red.), „The International Encyclopedia of Media Psychology" (Vol. 3). (The Wiley Blackwell-ICA International Encyclopedias of Communication). Wiley Blackwell, 2021.

Kolejnym elementem przemawiającym za wyborem tematu jest jego bezdyskusyjna przynależność tematu pracy do dyscypliny nauk o bezpieczeństwie. Definicję *deepfake* znaleźć można w licznych polskojęzycznych opracowaniach dyscypliny. Pojęcie zostało umieszczone między innymi w Vademecum bezpieczeństwa informacyjnego Olgi Wasiuty i Rafała Klepki[26]. Deepfake jako technologia niesie ze sobą nie tylko wysoki potencjał, lecz także bliżej niezidentyfikowane zagrożenia dla bezpieczeństwa państwa i jego obywateli. W toku prowadzonych badań zidentyfikowano je, opisano, a także zaproponowano schemat rozwiązań, mający za zadanie zminimalizować podatność społeczeństwa na fałszywe nagrania.

Za przynależnością tematu do dyscypliny przemawiają również użyte w niej metody badawcze, wykorzystywane w naukach o bezpieczeństwie. Są to przede wszystkim metody empiryczne – eksperyment oraz badanie literatury. Oprócz tego wykorzystano liczne metody teoretyczne, takie jak analiza, porównanie, abstrahowanie czy wnioskowanie.

Nadmienić należy, iż temat manipulacji rzeczywistościom przy wykorzystaniu technologii *deepfake* nie jest autorowi obcy. W ramach praktyk prowadził on zajęcia na Uniwersytecie Otwartym Uniwersytetu Warszawskiego pod tytułem „Deep fakes – czyli o tym, dlaczego w sieci lepiej pozostać niezauważonym" oraz „Deep Fakes – o krok od utraty wiary w rzeczywistość". Ponadto kilkukrotnie występował on na konferencjach zarówno ogólnopolskich jak i międzynarodowych, prezentując cząstkowe postępy niniejszych badań. Autor przeprowadził również kilkanaście szkoleń dotyczących dezinformacji i wpływu manipulacji obrazem na odbiorców. Autor swój warsztat związany z *deepfake* ulepsza praktycznie od momentu powstawania tej technologii. Tworzy nagrania wideo oraz manipuluje wizerunkami osób, tworzy nowe tożsamości i generuje przekłamane obrazy. Wszystkie te elementy świadczą o tym, iż temat nie jest autorowi obcy.

1.1.2 Przedmiot i cel badań

Dezinformacja multimedialna, w tym deepfake, to rodzaj fałszywej lub przekłamanej informacji, która jest sztucznie generowana lub modyfikowana za pomocą nowoczesnych technologii, takich jak uczenie maszynowe, sztuczna inteligencja, czy też

[26] O. Wasiuta, R. Klepka, „Vademecum bezpieczeństwa informacyjnego", 2020, t. 1 A-M, s. 266 – 270.

różnego rodzaju algorytmy. W przypadku deepfake jest to technologia pozwalająca na sztuczne generowanie wizerunku lub głosu człowieka w filmie, lub nagraniu, tak by przedstawiał on coś, co rzeczywiście nigdy nie miało miejsca. Technologia deepfake zdaje się być szczególnie niebezpieczna, ponieważ pozwala na tworzenie przekonujących fałszywych nagrań, które są trudne do odróżnienia od prawdziwych. Może to prowadzić do rozpowszechniania fałszywych informacji, które niosą za sobą poważne konsekwencje, zarówno na poziomie indywidualnym, jak i społecznym czy nawet państwowym.

Jeśli chodzi zaś o wpływ technologii deepfake na bezpieczeństwo narodowe, to jest ono potencjalnie bardzo wysokie. Fałszywe nagrania deepfake mogą być wykorzystywane do manipulowania opinią publiczną, osłabiania zaufania do rządów czy instytucji państwowych, a nawet siania chaosu i zamętu w kluczowych dziedzinach, takich jak polityka czy bezpieczeństwo wewnętrzne.

W związku z tym, badanie zjawiska dezinformacji multimedialnej opartej na deepfake jest ważne dla rozpoznania zagrożeń i znajdowania sposobów na ich zwalczanie. Dotyczy to zarówno rozwoju technologii i metod detekcji deepfake, jak i podejmowania działań edukacyjnych i komunikacyjnych, aby uświadomić ludziom, jakie zagrożenia płyną z takiej dezinformacji oraz jakie kroki należy podjąć w celu ochrony przed nią.

W pracy zidentyfikowane zostały dwa cele badawcze – cel główny oraz cel praktyczny (utylitarny). Celem głównym badawczym jest identyfikacja i analiza zagrożeń dla bezpieczeństwa narodowego, wynikających z rozwoju technologii deepfake. Istotą problemu jest poznanie i zrozumienie potencjalnych zagrożeń dla bezpieczeństwa państwa i jego obywateli, które mogą mieć swoje źródło w nadużyciach lub złych intencjach związanych z technologią deepfake. Eksperyment badawczy, przeprowadzony w niniejszej pracy, pozwala na identyfikację i ocenę ryzyka związanego z poszczególnymi zagrożeniami, które mogą mieć miejsce w różnych obszarach, takich jak gospodarka, cyberbezpieczeństwo, czy też działalność nielegalna lub terrorystyczna. W szczególności badanie pozwala na zrozumienie, jakie zagrożenia mogą wynikać z technologii deepfake.

Celem utylitarnym jest zaś wskazanie kierunków działania w celu ochrony przed dezinformacją. Polega on na zaproponowaniu konkretnych kroków i rozwiązań, które mogą być podejmowane w celu zniwelowania i zmniejszenia ryzyka zagrożenia, ale

również skuteczniejszej ochrony przed dezinformacją deepfake oraz zwiększania odporności społeczeństwa na tego rodzaju działania. Jest to cel istotny zarówno dla rządów, jak i dla przedsiębiorstw oraz organizacji, które mogą być potencjalnie narażone na działania dezinformacyjne.

Praca z charakteru badawcza, dzięki wyznaczeniu celu utylitarnego będzie miała wymiar praktyczny. Wyniki eksperymentu bezpośrednio przełożą się na wyznaczenie elementów pozwalających na identyfikację fałszu. Pozwoli to na zaproponowanie utworzenia standaryzowanego pakietu edukacyjnego obywatelskich umiejętności, wiedzy i kompetencji w zakresie rozpoznawania użycia technologii deepfake.

1.1.3 Problemy i zmienne w procesie badawczym

Najważniejszym elementem pracy badawczej jest sformułowanie problemu badawczego oraz próba odpowiedzi na niniejsze pytania, czyli przedstawienie hipotez. Problem badawczy rozumie się jako badanie relacji zachodzących pomiędzy zmiennymi – zmienną zależną – Y oraz zmienną niezależną – X. Właściwie postawione pytanie dotyczy, więc tego, czy dana zmienna X rzeczywiście wpływa na Y oraz jeżeli tak to, w jaki sposób[27]. W niniejszej pracy zmienną zależną nazywa się bezpieczeństwo narodowe, zaś zmienną niezależną dezinformację tworzoną przy wykorzystaniu technologii *deepfake*. Badany jest więc wpływ nagrań na owo bezpieczeństwo oraz elementy, które potęgują lub obniżają wspomniane efekty działania manipulujących nagrań deepfake.

Stefan Nowak wskazuje, że wyłącznie prawidłowo postawione pytanie badawcze (lub zbiór pytań) pozwala na przeprowadzenie badania oraz dalszą refleksję nad jego wynikami[28]. Pierwsze próby sformułowania pytań badawczych naszły autora w trakcie pierwszego roku szkoły doktorskiej. Zgłębiając literaturę dotyczącą dezinformacji, natrafiono na obszerne artykuły i monografie, opisujące „*fake news*-y" oraz wojnę informacyjną. Na drodze żmudnych dociekań teoretycznych odkryta została przestrzeń dezinformacji audiowizualnej, znacznie uboższa, w porównaniu do bogatej w historię dezinformacji treścią językową. Badania eksploracyjne pozwoliły ustalić, iż nowy w swoim pomyśle *deepfake* stanowi nieeksplorowaną dotychczas przez naukowców przestrzeń.

27 J. Brzeziński, „Metodologia badań psychologicznych", Warszawa 2004, s. 216.
28 S. Nowak, „Metodologia badań społecznych", Warszawa 2010, s. 214.

Według klasyfikacji pytań zaproponowanej przez Jerzego Giedymina klasyfikuje się pytania na pytania rozstrzygnięcia (czy?) i pytania dopełnienia (jak?)[29]. Problem badawczy można sformułować w sposób następujący, podając go w postaci zdania pytającego „Czy technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu?". Rozwiązanie problemu ma na etapie heurystycznym doprowadzić do rekonstrukcji przebiegu manipulacji nagraniami deepfake oraz sposobu jego oddziaływania na jednostkę.

Próbą odpowiedzi na to pytanie jest sześć hipotez szczegółowych, które pokrótce przedstawiono we wstępie, a które szczegółowo zostały opisane w następnym podrozdziale.

W celu realizacji zamierzenia badawczego należy podjąć się opracowania następujących zagadnień i odpowiedzieć na szereg następujących pytań, które z racji na swój charakter określone zostały pytaniami pomocniczymi.

• *Jaki jest status ontologiczny wiedzy na temat wpływu zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe?* (rozdział 2),

• *w jaki sposób powstają nagrania deepfake?* (rozdział 3),

• *Czy internauci rozróżniają materiały multimedialne prawdziwe od fałszywych?* (rozdział 4),

• *Jaki wpływ ma percepcja zmanipulowanych nagrań wideo na wybory i opinie internautów, w świetle bezpieczeństwa narodowego?* (rozdział 5),

• *Jaki wpływ mają zastosowane moderatory na odbiór dezinformacji deepfake, a przez to na bezpieczeństwo narodowe?* (rozdział 6),

• *Jakie należy podjąć działania w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake?* (rozdział 6).

Pytania te, sumarycznie, pomogą udzielić odpowiedzi na główne pytanie badawcze oraz zweryfikować stawiane hipotezy. Pytania 1 – 6 pozwolą również na udzielenie odpowiedzi na ostatnie pytanie, którego odpowiedź opierać się będzie na holistycznej analizie stawianych problemów.

29 J. Giedymin, „Problemy, założenia, rozstrzygnięcia: studia nad logicznymi podstawami nauk społecznych", Poznań 1964.

1.1.4 Hipotezy badawcze

O hipotezach w nauce o bezpieczeństwie pisali między innymi Tadeusz Jemioło oraz Andrzej Dawidczyk. Stwierdzają oni, iż hipoteza badawcza to „stwierdzenie, co, do którego istnieje pewne prawdopodobieństwo, że stanowi prawdziwe rozwiązanie badanego problemu". Dalej autorzy zaznaczają, iż hipoteza powinna zostać sformułowana w formie twierdzącej, a zatem nie powinna mieć charakteru zdania przeczącego, oceniającego, pytającego lub postulującego[30]. Hipoteza badawcza postawiona w niniejszej pracy brzmi: technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu. Badacze dopuszczają również zaproponowanie w pracy hipotez roboczych, zwanych niekiedy również hipotezami szczegółowymi. Mają one odpowiadać na problemy szczegółowe, jeżeli takie zostały w pracy uwzględnione. Zdanie to podziela między innymi Jacek Piwowarski[31].

Pierwsza z zaproponowanych hipotez szczegółowych dotyczy sytuacji zastanej, a jej weryfikacja opiera się na podstawie analizy literatury. Brzmi ona następująco: zmanipulowane materiały audiowizualne mogą mieć istotny wpływ na bezpieczeństwo narodowe, a obecny stan wiedzy na ten temat jest niepełny i nadal rozwijający się. Przebieg badań zmierzających do weryfikacji tej hipotezy zaprezentowano w rozdziale drugim.

Druga z roboczych hipotez dotyczy bezpośrednio możliwości tworzenia nagrań deepfake i tego, do czego są one wykorzystywane. Zawarta została w następującym twierdzeniu: nagrania deepfake powstają poprzez wykorzystanie oprogramowania do zmiany obrazu lub dźwięku w oryginalnym nagraniu, tak, aby wyglądało to jakby ktoś inny mówił lub wyglądał inaczej niż w rzeczywistości. Przebieg badań zmierzających do zweryfikowania tej hipotezy zaprezentowano w rozdziale trzecim, przy pomocy badań terenowych i przeglądu literatury.

Kolejna hipoteza szczegółowa dotyczy możliwości percepcji nagrań deepfake przez internautów. Brzmi ona: internauci częściowo rozróżniają materiały multimedialne prawdziwe od fałszywych, wytworzonych przy wykorzystaniu technologii deepfake, ale mogą mieć trudności z odróżnieniem prawdziwych materiałów od fałszywych,

[30] T. Jemioło, A. Dawidczyk, „Wprowadzenie do metodologii badań bezpieczeństwa", Warszawa 2008, s. 31–33.
[31] J. A. Piwowarski, „Metodologiczne i badawcze założenia pracy dyplomowej z dyscypliny nauk o bezpieczeństwie – przykład", [w:] Security, Economy & Law Nr 4/2019 (XXV), s. 32-34.

zwłaszcza, jeśli są one dobrze wykonane. Proces weryfikacji hipotezy opisany został w czwartym rozdziale i jest pochodną części wyników otrzymanych z eksperymentu badawczego.

Czwarta robocza hipoteza porusza temat wpływu nagrań deepfake na decyzje i zachowania internautów. Przebieg badań zmierzających do jej weryfikacji zaprezentowano w rozdziale piątym, na podstawie analizy otrzymanych w eksperymencie badawczym odpowiedzi. Treść hipotezy to: percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów, a jej oddziaływanie na bezpieczeństwo narodowe jest zależne od stopnia wiarygodności tych nagrań oraz od kontekstu, w jakim są one prezentowane.

Odpowiedzią na piąte pytanie badawcze, ma być niniejsza hipoteza: osoby z podwyższonymi poszczególnymi wskaźnikami społecznymi są bardziej podatne na uleganie manipulacjom dezinformacji deepfake, co przyczynia się do zwiększenia ryzyka zagrożenia bezpieczeństwa narodowego. Weryfikacja hipotezy przeprowadzona została w rozdziale szóstym, na podstawie analizy wyników badania, zestawionego z zastosowanymi moderatorami.

Analiza odpowiedzi na ostatnie pytanie badawcze, dotyczące modelu kontrdezinformacyjnego, została opisana w szóstym rozdziale. Hipotezą weryfikującą pytanie jest niniejsze zdanie oznajmiające: aby ochronić się przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, należy podjąć działania zapobiegające rozprzestrzenianiu się takich materiałów, edukować społeczeństwo w zakresie rozpoznawania dezinformacji audiowizualnej oraz zapewnić odpowiednie narzędzia do weryfikacji prawdziwości takich materiałów.

1.2 Metody badawcze

W poniższym podrozdziale opisane zostały zastosowane w pracy metody badawcze. W kolejności są to wpierw teoretyczne metody badawcze, opisujące procesy myślowe zachodzące w trakcie przetwarzania materiałów badawczych, następnie wymienione zostały metody empiryczne. Szczegółowo opisano metody najszerzej zastosowane w pracy, takie jak analiza dokumentów oraz eksperyment badawczy, poświęcając im osobne podrozdziały. To właśnie badanie w schemacie eksperymentalnym wewnątrz osobowym jest najważniejszym elementem niniejszej publikacji. Z tego powodu poświęcono mu najwięcej uwagi.

Na wstępie należy wytłumaczyć pojęcie „metody", które to wykorzystywane jest wielokrotnie w niniejszym podrozdziale. Metoda, z języka greckiego *methodos*, czyli badanie, oznacza sposób badania rzeczy i zjawisk, reguły dochodzenia do prawdy, określone normy badania rzeczywistości. Pojęcie *metody* badawczej definiowane jest z różnych punktów widzenia.

Janusz Sztumski rozumie metodę badawczą jako system założeń i reguł, które pozwalają na uporządkowanie empirycznej lub teoretycznej działalności naukowej, celem świadomego osiągnięcia zakładanego celu[32]. J. Sztumski zwraca uwagę na trzy elementy właściwe metodzie badawczej: wzorcowe wykonywanie badań naukowych, ich pisarskie opracowanie oraz krytyczna ocena.

Andrzej Czupryński definiuje metodę jako schemat postępowania, który jest określony, powtarzalny i wyuczony, świadomie stosowany do osiągnięcia określonego celu poprzez dobór odpowiednich środków. A. Czupryński wskazuje, iż niezmiernie ważnym elementem metody jest jej powtarzalność, pojawiająca się w każdej metodzie, z wyjątkiem metod heurystycznych[33].

Mieczysław Łobocki, przedstawiciel szkoły pedagogicznej, owo pojęcie tłumaczy, jako system reguł czy też inaczej szereg operacji poznawczych i praktycznych skierowanych na z góry założony cel badawczy. Mieści się w tym również kolejność oraz sposób użycia specjalnych środków i działań zmierzających do osiągnięcia celu[34].

Definicję nauk psychologicznych najszerzej przedstawia Józef Pieter. Jego zdaniem metody badawcze to wszelkie procesy, które zachodzą w trakcie badań od momentu powstania problemu do jego jakościowego i ilościowego opracowania wyników[35].

Jednak doszukując się genezy istnienia pojęcia metod badawczych w naukach społecznych, należy spojrzeć do prekursora metodologicznego – Stefana Nowaka. Według jego prostej definicji pod pojęciem tym kryje się powtarzalny i skuteczny sposób rozwiązywania ogólnego problemu badawczego[36].

---

[32] J. Sztumski, „Wstęp do metod i technik badań społecznych", Katowice 2005, s. 65 – 70.
[33] A. Czupryński, „Metoda naukowa", [w:] Nauki o bezpieczeństwie. Wybrane problemy badań.", red. A. Czupryński, B. Wiśniecki, J. Zboina, Józefów 2017, s. 17.
[34] M. Łobocki, „Metody badań pedagogicznych", Warszawa 1982, s. 112 – 116.
[35] J. Pieter, „Ogólna metodologia pracy naukowej", Wrocław 1967, s. 70.
[36] S. Nowak (red.), „Metody badań socjologicznych. Wybór tekstów", Warszawa 1965, s. 12 – 13.

1.2.1 Teoretyczne metody badawcze

W procesie badawczym wykorzystywane zostały zarówno teoretyczne, jak i empiryczne metody badawcze. Różnią się one zarówno swoim charakterem, jak i przeznaczeniem. O ile metody empiryczne kładą nacisk na zebranie materiału badawczego, tak metody teoretyczne mają za zadanie uporządkować i odpowiednio go przetworzyć. Metody te mają charakter opisowy i przedstawić je można w postaci schematów logicznych. Piotr Sienkiewicz nazywa je metodami ścisłego wnioskowania lub inaczej ścisłego rozumowania[37].

Szeroki katalog metod oraz ich dokładny opis zaprezentowany został przez Zenona Chojnackiego i Mariana Cieślarczyka[38]. Dokonali oni dogłębnej analizy literatury przedmiotu, wyróżniając i opisując szczegółowo metody badawcze teoretyczne.

1. Analiza

Analiza u Z. Chojnackiego i M. Cieślarczyka stanowi metodę z najbogatszą podgrupą metod badawczych teoretycznych. W jej skład naukowcy zaliczają takie metody badań formalnych jak analiza ilościowa, jakościowa, logiczna czy przyczynowa, a także krytyka źródeł i piśmiennictwa. Dalej analiza w ocenie autorów może być elementarna lub strukturalna, funkcjonalna, pojęciowa, porównawcza czy genetyczna lub matematyczna. Autorzy wyróżniają również analizy metod mieszanych, takich jak analizy jakościowo – ilościowe, systemowe czy wartości. W ich ocenie metoda badawcza analizy to intelektualny podział części składowych obiektu badań w celu wyróżnienia i zbadania ich cech, a także wspólnych relacji ich łączących[39].

2. Abstrahowanie

Abstrahowanie ma na celu wyodrębnienie kluczowych dla badacza elementów składowych badanego podmiotu. Odbywa się to poprzez eliminację elementów drugorzędnych w danym badaniu lub zmniejszenie priorytetu ich wagi i poddanie ich analizie w dalszej kolejności. Wyróżnia się dwa rodzaje abstrahowania – izolujące i generalizujące. Abstrahowanie izolujące ma na celu wyizolowanie danego elementu badania od innych obiektów i poddanie go analizie lub syntezie. Abstrahowanie

[37] P. Sienkiewicz, „Metody badań nad bezpieczeństwem i obronnością", Warszawa 2010, s. 35 – 38.
[38] M. Cieślarczyk, Z. Chojnacki, M. Mróz, S. Sirko, „Metody techniki i narzędzia badawcze oraz elementy statystyki stosowane w pracach magisterskich i doktorskich", Warszawa 2006, s. 45 – 60.
[39] Tamże.

generalizujące ma zaś zmierzać do przeprowadzenia uogólnienia na bazie wielu elementów wschodzących w skład danego badanego zbioru. Ma to na celu sformułowanie ogólnych zasad, reguł, stosunków panujących w danym obszarze badawczym.

3. Porównanie

Porównanie to intelektualne identyfikowanie różnic oraz podobieństw badanym obiekcie, w odniesieniu do innych obiektów, zbliżonych swoimi właściwościami. Inaczej mówiąc, jest to akcja lub proces zestawienia ze sobą dwóch lub więcej rzeczy, lub pojęć, w celu określenia ich podobieństw i różnic. Może to obejmować analizowanie ich właściwości, cech, działania lub innych aspektów. Porównanie może być używane do wyciągania wniosków, podejmowania decyzji lub do celów edukacyjnych.

4. Uogólnienie

Uogólnienie, inaczej nazywane generalizacją, ma na celu sformułowanie ogólnych twierdzeń, dotyczących badanego problemu, zjawiska. Odbywa się to na podstawie badania częściowego wycinka rzeczywistości, pojedynczego zjawiska lub mniej złożonego od obiektu będącego celem generalizacji. W naukach o bezpieczeństwie wyróżnia się dwa rodzaje uogólnienia: historyczne sprawozdawcze oraz historyczne indukcyjne. Twierdzenia pierwszego nie wykraczają poza przebadane przypadki. Przy drugim zaś – uogólnieniu historycznym indukcyjnym – powszechnym jest wykraczanie poza przebadany materiał. Odnosi się to również do przypadków przeszłych, których analiza i badanie może dać nam wiedzę o przewidywaniach przypadków przyszłych.

5. Klasyfikowanie

Klasyfikowanie ma na celu uporządkowanie zebranego materiału badawczego. Na potrzeby klasyfikacji tworzy się odpowiednie kryteria i zasady, następnie według których kataloguje się zebrane elementy. Grupy elementów posiadających lub nieposiadających danej cechy nazywane są podzbiorami, a przy ich tworzeniu badacz kieruje się zasadami poprawności logicznej.

6. Wnioskowanie

Wnioskowanie odbywa się przy pomocy zasad logiki formalnej i pozwala ze znanych już twierdzeń wyprowadzać nowe. Przy wykorzystaniu procesu myślowego – rozumowania – badacz określa jedno lub kilka twierdzeń, uznaje je za prawdziwe, a następnie na ich podstawie dowodzi prawdziwości nowego twierdzenia. Wyróżnia się wnioskowanie dedukcyjne, indukcyjne, redukcyjne oraz odbywające się poprzez analogię.

Wnioskowanie dedukcyjne, inaczej nazywane jest niezawodnym i polega na wyprowadzaniu nowych twierdzeń na podstawie pierwotnych przesłanek, wcześniej uznanych za prawdziwe, lecz znajdujących swą poprawność uzasadnienia wyłącznie z zastosowaniem rozumowania logicznego.

Wnioskowanie indukcyjne określa się uprawdopodobniającym, ponieważ wysnuwanie nowych twierdzeń odbywa się za pomocą obserwacji znanych faktów jednostkowych. Poprzez tłumaczenie uogólniające, z pomocą wnioskowania upodobniającego, formułuje się nowe wnioski ogólne. Szczegółowa przesłanka badanego zjawiska pozwala zatem dane zjawisko przypisać całości.

Przeciwstawnym dla wnioskowania dedukcyjnego, jest wnioskowanie redukcyjne, zwane również inwersyjnym. Proces zmierza do wyprowadzenia nowych twierdzeń poprzez dobieranie do niego takich przesłanek, z których owo twierdzenie logicznie wynika (na zasadach logiki formalnej). W tym przypadku to następstwa wnioskują o racjach. Nie należy mieć jednak pewności co do tego czy wniosek wysnuty przy pomocy wnioskowania redukcyjnego jest prawdziwy, co wynika z niezgodnością takiego rozumowania z kierunkiem wynikania logicznego.

Ostatnim elementem tej podkategorii jest wnioskowanie poprzez analogię. Metoda ta mówi o tym, że skoro dwa badane zjawiska czy obiekty są do siebie pod pewnymi względami podobne, to mogą być również podobne i w innych aspektach, nawet jeśli dana cecha została wykryta wyłącznie w jednym obiekcie czy zjawisku. Należy nadmienić, że zazwyczaj porównaniu podobieństwa podlegają najistotniejsze elementy.

7. Synteza

Metodą badawczą, mającą na celu intelektualne łączenie badanych elementów danego obiektu i sformułowanie dla niego całościowych wyników jest synteza. Poprzedzona powinna być dogłębną analizą badanego obiektu oraz jego elementów. Zdarza się, że synteza przyjmuje charakter uogólnienia porównawczego, jednak zazwyczaj stanowi odrębną metodę badawczą, zwierającą w sobie metody abstrahowania, porównania, grupowania czy uogólniania oraz wnioskowania.

1.2.2 Empiryczne metody badań, techniki i narzędzia badawcze

W ostatnim etapie pracy wykorzystane zostały rezultaty eksperymentu badawczego. Kilku ekspertom z zakresu zwalczania dezinformacji zostały ukazane

wstępne wyniki. Następnie poproszeni zostali oni o komentarz i wskazanie w oparciu o arkusze potencjalnych słabych i silnych stron poszczególnych metod dezinformacji obrazem oraz eksplikację własnych scenariuszy dalszego rozwoju zagrożenia.

1.2.2.1 Metoda badawcza – analiza dokumentów

Celem wstępnego rozpoznania problemu przeprowadzona została analiza danych zastanych (*retrospective data analysis*), zwana również badaniami zza biurka (*desk research*) lub badaniami wtórnymi (*secondary data analysis*) na podstawie studium szeroko pojętej literatury przedmiotu (analiza przeprowadzonych eksperymentów, dotychczasowych wyników badań, obserwowanych studiów przypadków, analiz i prognoz w zakresie rozwoju technologii *deepfake*, tzw. szarej literatury[40] oraz klasycznie pojmowanego piśmiennictwa stricte akademickiego). Synteza wniosków została sporządzona i opisana w niniejszej pracy. Analiza danych zastanych, zdaniem części badaczy[41], należy do technik analitycznych stosowanych w badaniach niereaktywnych (*non-reactive research*) określanych również badaniami niejawnymi (*unobtrusive research*)[42]. Inaczej mówiąc, jest to rodzaj badań, w którym analizuje się dane z przeszłości, które zostały już zgromadzone w celu odpowiedzi na pytanie badawcze lub do rozwiązania problemu. Badacz pod żadnym pozorem nie ingeruje w rzeczywisty obiekt analizy bądź charakter analizowanego zjawiska. W przeciwieństwie do badań prospektywnych, w których dane są zbierane specjalnie dla konkretnego celu badawczego, analiza danych zastanych polega na wykorzystaniu danych już istniejących, które nie były zbierane w celu rozwiązania konkretnego problemu badawczego. Do zalet tej metody możemy zaliczyć przede wszystkim brak wpływu na przedmiot badania, łatwą dostępność i niskie koszty[43]. Ponieważ informacje pochodzą z dokumentów i zasobów dostępnych w Internecie oraz z bibliotek, archiwów i innych źródeł danych, nie wymagają bezpośredniego kontaktu z badanymi lub eksperymentów terenowych. Wymagają one jednak zwykle posiadania dostępu do różnych źródeł informacji oraz umiejętności ich właściwego przetwarzania[44]. Należy

[40] Poprzez pojęcie „szarej literatury" rozumieć należy publikacje trudno dostępne, niepublikowane i nierecenzowane przez niezależnych recenzentów.
[41] E. Babbie, „Podstawy badań społecznych", Wydawnictwo Naukowe PWN, Warszawa 2013.
[42] E. J. Webb, D. T. Campbell, R. D. Schwartz, L. Sechrest, J. B. Grove, „Nonreactive Measures in the Social Sciences", Dallas: Houghton Mifflin, 1981.
[43] E. Babbie, „Badania społeczne w praktyce", Warszawa 2004: Wydawnictwo Naukowe PWN. s. 341.
[44] L. Hofferth, „Secondary Data Analysis in Family Research", Journal of Marriage and Family, 67(4), 2005, s. 893.

jednak zaznaczyć, że badania zza biurka mają pewne ograniczenia, takie jak brak kontroli nad warunkami zbierania danych, trudności w uzyskaniu dostępu do wszystkich istotnych informacji, czy też brak możliwości uzyskania danych jakościowych (np. opinie, komentarze). Niewątpliwą wadą metody analiz danych zastanych jest również potencjalna nieaktualność danych, ich nieadekwatność dla problemu badawczego będącego przedmiotem zainteresowania badacza, konieczność doskonałej znajomości technik eksploracji danych, wyczucie źródłoznawcze, a także wiedza o metodologii badań, bowiem ewaluacji ze strony badacza podlegają nie tylko wyniki badań, lecz także, co niemniej ważne, procedury ich zgodnego ze standardem pozyskiwania.

W niniejszym przypadku zaletą tej metody jest rozległość badanych obszarów. Konieczne było przeanalizowanie literatury technicznej, z zakresu tworzenia deepfake jak i też z zakresu psychologii społecznej oraz socjologii, celem prawidłowego przygotowania i przeprowadzenia eksperymentu. Atutem jest również bogactwo danych, na które natrafiono[45]. Wartość poznawcza i heurystyczna tej metody szczególnie wzrosła wraz z uczynieniem Internetu medium masowym i powstaniem zjawiska Big Data[46], konsekwentnie dzięki coraz większej liczbie publikacji w wolnym dostępie (*open access*).

1.2.2.2 Metoda badawcza – eksperyment badawczy

Eksperyment badawczy to metoda pozwalająca na przeprowadzenie doświadczenia naukowego w celu sprawdzenia hipotez lub teorii. Eksperyment polega na manipulowaniu jedną lub większą liczbą zmiennych nazywanych zmiennymi eksperymentalnymi, a następnie ocenie, czy zmiana tych zmiennych wpływa na inne zmienne, zwane zmiennymi objaśnianymi.

Celem eksperymentu jest zdobycie bezpośrednich dowodów na poparcie lub obalenie hipotezy, co pozwala na rozwijanie lub modyfikowanie obecnych teorii naukowych. To jedno z podstawowych narzędzi badawczych stosowanych w naukach przyrodniczych i społecznych, a także w innych dziedzinach, takich jak psychologia, ekonomia czy medycyna.

[45] Z. Bednarowska, „Desk research – wykorzystanie potencjału danych zastanych w prowadzeniu badań marketingowych i społecznych", Marketing i Rynek, 7, 2015, s. 18-26.
[46] M. P. Johnston, „Secondary Data Analysis: A Method of which the Time Has Come. Qualitative and Quantitative Methods in Libraries, 2014, 3, 2014, s. 619-626.

W Polsce, w naukach społecznych jako jeden z pierwszych eksperyment społeczny kompleksowo opisał Antoni Sułek[47]. Definiuje on eksperyment jako „powtarzalny zabieg polegający na planowej zmianie przez badacza jednych czynników w badanej sytuacji, przy równoczesnej kontroli innych czynników, podjęty w celu uzyskania w drodze obserwacji odpowiedzi na pytanie o skutki tej zmiany”[48].

Istnieją dwa podstawowe rodzaje eksperymentów: eksperymenty laboratoryjne i eksperymenty naturalne. Eksperymenty laboratoryjne przeprowadzane są w warunkach kontrolowanych, co pozwala na precyzyjne manipulowanie zmiennymi eksperymentalnymi i dokładne monitorowanie ich wpływu na zmienne objaśniane. Eksperymenty naturalne przeprowadzane są w warunkach naturalnych i polegają na obserwacji zjawisk zachodzących w przyrodzie.

Przeprowadzenie wiarygodnego eksperymentu wymaga zadbania o odpowiednią kontrolę warunków oraz zastosowanie metod losowania do przydzielania badanych do poszczególnych grup. Dzięki temu możliwe jest zminimalizowanie wpływu czynników zewnętrznych na wyniki, co zwiększa ich wiarygodność.

1.2.2.2.1 Ograniczenia badawcze w eksperymencie

W naukach społecznych spotkać można tendencję do rozbudowywania tytułów prac naukowych. Prawdopodobnie dzieje się tak z powodu obawy o zbyt szerokie ujęcie opisywanego problemu. Podejście to odstaje od „zachodniego” modelu nauki, gdzie zdając sobie sprawę z mnogości źródeł, dąży się w sposób najprostszy do syntetycznego prezentowania wyników. W tym akapicie umieszczone zostały wszelkie ograniczenia badawcze (ograniczniki), jakie przyjął autor na potrzeby niniejszych badań, a jakie nie zostały ujęte w tytule pracy. Zgodnie z przyjętą metodologią, założone zostały następujące ograniczenia przyjęte w procesie badawczym: czasowe, przedmiotowe, warsztatowe, technologiczne oraz źródłowe[49].

- Ogranicznik czasowy – badany okres czasowy ograniczony jest datami powstania technologii deepfake – listopad 2017 rok oraz datą zakończenia badań empirycznych – czerwiec 2022 rok. Tym samym ujęty okres obejmuje cztery lata, w trakcie których technologia deepfake nabierała popularności i była doskonalona w swoim działaniu.

[47] A. Sułek, „Eksperyment w badaniach społecznych”, Warszawa 1979.
[48] Tamże, s.15.
[49] T. Pawłuszko, „Wstęp do metodologii badań politologicznych”, Częstochowa 2013, s. 12.

- Ogranicznik przedmiotowy – w pracy nie został podjęty temat manipulacji obrazem w zależności od czasu wykorzystania. Chociaż jest on niezwykle ważny, w ocenie autora, z racji swojej objętości, nie sposób było zbadać również i ten obszar.
- Ogranicznik warsztatowy – w niniejszej dysertacji pominięte zostały aspekty prawne podejmowanego tematu. Autor w swojej ocenie nie czuje się kompetentny do prowadzenia badań w tym aspekcie, z racji braku wykształcenia prawniczego.

  Ponadto badaniu w schemacie eksperymentalnym wewnątrz osobowym, poddane zostały wyłącznie osoby z grupy wiekowej 18 – 33 lata.

  Wybranie tej grupy społecznej uzasadnia się pilotażowym charakterem badania, ograniczonym budżetem, łatwością dotarcia do osób badanych, a także chęcią zbadania osób aktywnych w Internecie. Studenci, z racji swojego wieku oraz doświadczenia, są potencjalnie najbardziej świadomi postępującej dezinformacji oraz tym samym potencjalnie najbardziej na nią odporni.

  Ważnym ograniczeniem warsztatowym było również przygotowanie nagrań deepfake z polskimi influencerami mediów społecznościowych, którzy nie są aż tak rozpoznawani przez społeczeństwo jak czołowi polscy celebryci czy gwiazdy telewizyjne. Konieczne było jednak zminimalizowanie ryzyka wycieku nagrań oraz jego potencjalnych konsekwencji. Wstępny pomysł przygotowania nagrań deepfake z udziałem polityków mógłby zostać wykorzystany w celach politycznych i w negatywny sposób przyczynić się do rozwoju pracy.

  Przy odtworzeniu poniższego badania, zaleca się zorganizowanie go przy świadomej zgodzie osób, których wizerunek jest wykorzystywany oraz przy wsparciu instytucji badawczej organizującej badanie.
- Ograniczniki źródłowe – w pracy pominięta została bogata literatura chińska. Wynika to z bariery językowej, jakiej nie był w stanie pokonać autor, również ze wsparciem automatycznych słowników i internetowych tłumaczy.
- Ograniczniki technologiczne – planowane było użycie w badaniu okulografu (Eye tracking), mającego śledzić wzrok osoby badanej, by tym

samym wskazać elementy przyciągające i dominujące uwagę widza. Niestety ze względu na obostrzenia pandemiczne obowiązujące w trakcie badań, zrezygnowano z tej dodatkowej formy testowania.

1.2.2.2.2 Obszar, przebieg i organizacja eksperymentu

Z uwagi na stan epidemii, panujący w Polsce, w trakcie trwania badania, próba pozyskania respondentów okazała się dużym wyzwaniem. We wstępnym etapie planowano prowadzić badanie w warunkach laboratoryjnych, tj. sali komputerowej, gdzie osoby badane miałyby jednakowe warunki oraz zapewnione wsparcie techniczne. Za przeprowadzeniem badania w warunkach laboratoryjnych przemawiał również dodatkowy argument, jakim jest bezpieczeństwo wytworzonych na potrzeby badania materiałów. Zamknięte środowisko wyeliminowałoby wówczas całkowicie ryzyko wykradzenia i opublikowania nagrań przez osobę badaną. Jednakże, ze względu na ograniczenia covidowe, na dalszym etapie pracy zrezygnowano z tego pomysłu. Zgodnie z zaleceniem Dziekana Wydziału Bezpieczeństwa, Logistyki i Zarządzania WAT[50] studenci uczelni rok 2021 / 2022 mieli rokiem nauki zdalnej, co tyczyło się również zajęć w salach komputerowych. Tym samym, nieobecność studentów na uczelni poważnie utrudniała możliwość pozyskania uczestników eksperymentu.

W związku z tym zdecydowano się na przeprowadzenie eksperymentu w formie zdalnej. Celem maksymalnego zabezpieczenia nagrań oraz pozostałych materiałów, zdecydowano się na samodzielne utworzenie platformy badawczej, na której następnie umieszczono badanie.

Wyłączono możliwość pobierania nagrań, ukryto również ścieżkę do nich prowadzącą. Utworzono indywidualne konta dla uczestników badania, w liczbie 600 oraz zabezpieczono je indywidualnym hasłem. Zablokowano również możliwość ponownego zalogowania się na platformę przy użyciu tego samego loginu i hasła. Wszystkie te działania przeprowadzono z myślą o zachowaniu maksymalnej ostrożności i bezpieczeństwa.

Proces przygotowywania materiałów, ze względu na swój wrażliwy charakter, przeprowadzony został w odizolowanym środowisku, w warunkach laboratoryjnych. Całość badania hostowana była na prywatnym serwerze z ograniczonym dostępem.

[50] Strona główna Wydziału Bezpieczeństwa, Logistyki i Zarządzania WAT, zalecenie Dziekana WBLiZ, https://wlo.wat.edu.pl/wp-content/uploads/2020/04/wytyczne.pdf [dostęp: 01.12.2022].

Żaden uczestnik nie miał kontaktu z innym badanym, uczestnicy zaakceptowali również klauzule do zachowania tajemnicy. Liczba całkowita uczestników badania wyniosła ponad 100 osób, jednak tylko 82 ukończyły badanie na dopuszczalnym etapie (obejrzenie wszystkich sześciu filmów i udzielenie odpowiedzi na pytania).

Na potrzeby badania przygotowanych zostało sześć nagrań. Dwa z nich powstały przy pomocy technologii deepfake, gdzie na twarze dwójki mało znanych osób (kobieta i mężczyzna) nałożono wizerunki popularnych influencerów (milion i ponad dwa miliony obserwujących na portalu Instagram[51]). Proces tworzenia nagrań deepfake został szczegółowo opisany w rozdziale trzecim.

Kolejne dwa nagrania (dwóch mężczyzn) odtworzono z faktycznych nagrań dwójki mniej znanych influencerów (oba konta ok. 500 tys. obserwujących na Instagramie). Ostatnie dwa nagrania to filmy prezentujące dwie nieznane osoby (kobieta i mężczyzna). Nagrania te charakteryzowały się niską jakością gry aktorskiej oraz słabą jakością wideo. Te cztery nagrania wykorzystywane były pierwotnie do zachęcania oglądających do rejestracji na fałszywej platformie inwestycyjnej, natomiast treść wszystkich sześciu nagrań była do siebie merytorycznie zbliżona. Szczegółowy opis treści filmów prezentowany jest w podrozdziale 1.2.2.2.3.

We wstępnej części badania, prezentowany był tekst marketingowy, a następnie niezbędne zgody[52], wymagające akceptacji osoby biorącej udział w badaniu oraz informacja o badaniu[53].

W pierwszej części badania respondenci uzupełniali ankiety online, stanowiące moderatory badania. Decyzja o ich wyborze szczegółowo opisana została w kolejnym podrozdziale. Były to:

- moderator lęku – GAD 7 – 7 pytań,
- moderator na depresyjność – PHQ-9 – 9 pytań,
- moderator potrzeby poznawczego domknięcia – 15 pytań,
- moderator kodów moralnych – MFQ – 30 pytań,
- moderator samooceny – SES – 10 pytań.
- moderator impulsywności – BIS-Brief – 8 pytań.

[51] Instagram – fotograficzny serwis społecznościowy hostingu zdjęć, połączony z aplikacją o tej samej nazwie, który umożliwia użytkownikom edycję zdjęć i filmów, stosowanie do nich filtrów cyfrowych oraz udostępnianie ich w różnych serwisach społecznościowych. https://www.instagram.com/ [dostęp: 01.01.2023].
[52] Teksty dostępne w załączniku.
[53] Teksty dostępne w załączniku.

Następnie, w losowej kolejności, wyświetlane były przygotowane nagrania wideo. Po każdym wyświetleniu następowała seria piętnastu zamkniętych, obowiązkowych pytań oraz dwa pytania opisowe, fakultatywne[54]. Ankiety zostały ujednolicone. Wszystkie pozycje zostały ocenione na 10 – punktowej skali (1 = w ogóle, 10 = bardzo).

Po udzieleniu odpowiedzi na pytania do wszystkich sześciu filmów pojawiało się 11 pytań metryczkowych, dotyczących sytuacji osoby badanej. Była ona proszona o podanie swojej płci, wieku, charakteru miejsca zamieszkania, wykształcenia, sytuacji zawodowej, kierunku ukończonych / odbywanych studiów, zawodów rodziców, sytuacji finansowej zamieszkiwanego gospodarstwa domowego, stanu cywilnego, szacunkowego dziennego czasu spędzanego w Internecie oraz preferencji wyborczych[55].

W kolejnym kroku osoba badana otrzymywała odkłamanie badania, czyli treść informującą ją o rzeczywistym celu badania – sprawdzenia jej percepcji przygotowanych nagrań[56].

Następnie osoba uzupełniająca ankietę była proszona o udzielenie odpowiedzi na ostatnie siedem pytań, dotyczących bezpośrednio wiedzy o technologii deepfake[57]. Pytania dotyczyły świadomości istnienia nagrań deepfake, udostępniania fałszywych nagrań, natrafiania na zmanipulowane filmy, zarówno w mediach społecznościowych jak i ogólnie w Internecie, osobistych odczuć dotyczących rozpoznania nagrań deepfake w niniejszej ankiecie oraz w całej sieci. Ostatnie pytanie dotyczyło niepokoju związanego z rozwojem deepfake, na pięciopunktowej skali.

1.2.2.2.3 Treść nagrań deepfake oraz pozostałych filmów

Eksperyment zakładał wyświetlenie w losowej kolejności sześciu nagrań, każde trwające od 30 do 60 sekund i zawierające elementy wspólne – zachęcenie do inwestycji na fałszywej platformie inwestycyjnej.

W Internecie znaleziono 2 nagrania dwoje średnio rozpoznawanych influencerów (mających pomiędzy 500 a 600 tysięcy obserwujących je osób), którzy na owych filmach zachęcali na zasadach afiliacji do zainwestowania na oszukańczej platformie inwestycyjnej. Każde z nagrań zawierało charakterystyczne, wymienione poniżej

[54] Treść pytań dostępna w załączniku.
[55] Szczegółowa treść pytań metryczkowych dostępna w załączniku.
[56] Treść odkłamania dostępna w załączniku.
[57] Treść pytań dostępna w załączniku.

elementy. Wpierw był to krótki wstęp związany z działalnością internetową danej osoby, celem uwiarygodnienia jej postaci. Następnie pojawiło się przejście do tematu aktualnej sytuacji finansowej w Polsce. W pierwszym przypadku był to temat rosnącej inflacji, zaś w drugim, utrudniony proces oszczędzania pieniędzy w dobie pandemii COVID19. W dalszej części pojawiała się informacja, że warto zainteresować się kryptowalutami, jako walutą przyszłości. Influencer informował następnie, że ciężko jest zacząć, nie mając żadnej wiedzy, ale jest pewna platforma „XYZ" do inwestowania, z której sam korzysta, odnosząc duże sukcesy, zachęcając do jej wypróbowania. Na koniec pojawia się informacja, iż konsultanci platformy wszystko najlepiej wytłumaczą i nauczą szybkiego pomnażania pieniędzy, choćby nawet niskich kwot. Ostatnim elementem było zaproszenie do kliknięcia w link zamieszczony w opisie filmu oraz w opisie profilu.

Na tej podstawie zdecydowano się utworzyć dwa, analogicznie wyglądające nagrania deepfake, z bardziej znanymi osobami. Napisano dwa, zamieszczone poniżej scenariusze, które zaprezentowano dwojgu znajomym, by wypowiedzieli poniższe kwestie przed kamerą. Scenariusz pierwszy (podszywanie się pod średnio znanego influencera) brzmiał następująco:

„Siemka, ja właśnie wróciłem do domu po treningu, testuję nowe suplementy diety, ale nie o tym dzisiaj chciałem mówić. Są tematy ważniejsze i pilniejsze. Jak pewnie zdążyliście się zorientować, ceny w sklepach z dnia na dzień są coraz wyższe, inflacja szaleje, a oprocentowanie na lokatach praktycznie zerowe. Trudno jest cokolwiek zaoszczędzić, a co dopiero zyskać.

Każdemu, kto chce zarobić trochę grosza, polecam gorąco zainteresowanie się tematem kryptowalut. Oczywiście nie jest łatwo samemu wszystko zrozumieć, jednak są specjaliści, którzy za niewielką prowizję pokażą nam jak otrzymać duży zysk. Giełda "XYZ" to miejsce, gdzie sam inwestuję i dzięki pomocy osobistego konsultanta odnoszę spore sukcesy. Link do strony „XYZ" znajdziecie w moim opisie. Wystarczy podać w formularzu swoje imię i nazwisko oraz numer telefonu, a konsultant niezwłocznie oddzwoni do Ciebie z ofertą zysków. Platforma „XYZ" to naprawdę olbrzymia szansa dla każdego z nas do zmiany swojej przyszłości na lepsze!".

Scenariusz drugi (podszywanie się pod znaną influencerkę i aktorkę) brzmiał następująco:

„Witajcie kochani!

Przychodzę dzisiaj do Was z lokami w stylu Hailey Bieber, kto jeszcze nie widział, zapraszam do mojej najnowszej rolki. Ale dzisiaj mniej optymistycznie. Widzicie, co się dzieje w kraju, jak inflacja idzie w górę i wszystko jest coraz droższe. Bardzo ciężko jest oszczędzać. Mój dobry znajomy polecił mi ostatnio bardzo dobrą platformę inwestycyjną „XYZ", gdzie inwestować możecie również w kryptowaluty. Jest to naprawdę pieniądz przyszłości i uważam, że warto tam zainwestować. Ja już korzystam i odnoszę bardzo duże sukcesy w "XYZ". Polecam Wam, naprawdę nie ma się czego bać. Doradcy wytłumaczą wszystko krok po kroku, poprowadzą tak jakby za rączkę, tak więc będziecie mogli bez problemu pomnożyć swoje pieniądze. Tak więc kliknijcie w link w opisie i zarejestrujcie się jeszcze dzisiaj. Powodzenia!"

Ostatnie dwa nagrania, to nagrania również znalezione w Internecie, reklamującą tą samą fałszywą platformę inwestycyjną „XYZ", natomiast prezentowaną przez nieznane osoby, niemające doświadczenia w nagrywaniu filmów czy prezentacji własnego wizerunku. Oba pochodzą z wczesnego działania platformy „XYZ" i oba charakteryzują się słabą jakością wideo oraz dźwięku.

W obu nagraniach osoby w nich występujące pytały się, czy oglądający je wie, jak zarobić duże pieniądze, a następnie machając plikiem banknotów, informowały o istnieniu platformy „XYZ", na której w łatwy sposób można pomnożyć zainwestowane środki. Oba filmy były dużo krótsze od pozostałych czterech i mieściły się w czasie 30 sekund.

1.2.2.2.4 Procedura doboru uczestników

Eksperyment przeprowadzony został na grupie osób w wieku 18 – 33 lata. Każda osoba z grupy otrzymywała jednakowy zestaw pytań oraz identyczne nagrania. Kwestionariusze moderatorów, jak i filmy wyświetlały się każdemu w losowej kolejności.

Aby zmniejszyć prawdopodobieństwo braku danych z powodu wycofania się uczestników w trakcie badania (uczestnik badania w każdej chwili mógł z niego zrezygnować), starano się zrekrutować 100 osób. Po usunięciu przypadków, w których brakowało więcej niż 5% danych (n = x), analizie poddano odpowiedzi łącznie 82 uczestników – osób w wieku 18 – 33 lata. Średni wiek mężczyzn wyniósł 23 lata, zaś kobiet 24 lata. Zachowana została proporcja płci, procent przebadanych kobiet wyniósł 61%, zaś mężczyzn 39%. Kwestionariusz internetowy w trakcie rekrutacji wypełniło 90

osób. Ankieta internetowa rozsyłana była głównie do cywilnych studentów oraz absolwentów Wojskowej Akademii Technicznej oraz Uniwersytetu Warszawskiego. Uczestnicy wyrazili zgodę na udział w badaniu, którą wypełnili na początku i na końcu kwestionariusza. Z danych osobowych uczestnicy zostali poproszeni o podanie wyłącznie swojego wieku i płci oraz poziomu wykształcenia.

Nie zastosowano odchyleń badawczych. Wszyscy uczestnicy odpowiedzieli na wszystkie sześć ankiet moderujących, które miały zastosowanie do indywidualnych cech osobowości. Pozwoliło to wykorzystać każdego uczestnika jako własną kontrolę, porównując odpowiedzi do każdego scenariusza. Czas udzielania odpowiedzi przez respondentów wyniósł średnio około 37 minut.

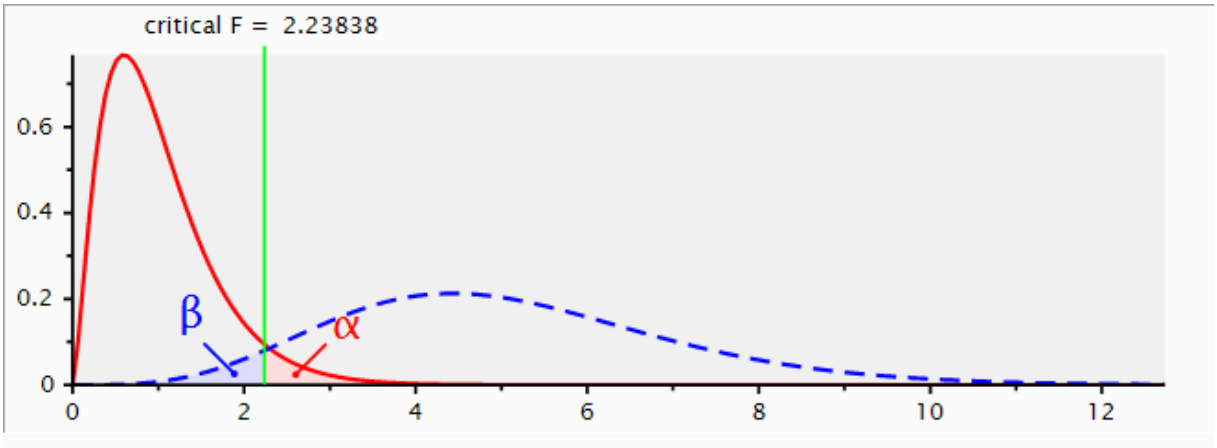
1.2.2.2.5 Obliczenie minimalnej próby

W celu obliczenia wymaganej liczebności próby badawczej postawiono następujące założenia[58]:

1. Do przeprowadzenia głównych analiz statystycznych zostanie wykorzystana analiza wariancji z powtarzanymi pomiarami (RM ANOVA),
2. W ramach analiz wariancji z powtarzalnymi pomiarami zastosowane zostaną porównania wewnątrzgrupowe (w zakładanym schemacie badawczym nie będzie porównań międzyosobniczych),
3. Każdy uczestnik badania będzie miał za zadanie obejrzeć sześć filmów – i do każdego z nich udzielić odpowiedzi na serię pytań,
4. Minimalna wielkość efektu, który planuje się zaobserwować, wynosi 0,15,
5. Błąd pomiaru ustalono na ogólnoprzyjętym poziomie α =0,05,
6. Moc testu (1 – β) ma wynosić 0,95.

Na podstawie powyższych założeń, obliczono (w programie G*Power 3.1.9.7[59]), że grupa badawcza musi się składać z minimum 75 osób badanych[60]. Pełen protokół z analizy znajduje się w załączniku nr 1, poniżej zaś znajduje się wykres wygenerowany w programie G*Power.

[58] Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, s. 175-191.
[59] Aplikację G*POWER pobrano ze strony producenta aplikacji – https://g-power.apponic.com/ [dostęp: 28.01.2023].
[60] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, s. 1149-1160.



*Wykres 1 Print Screen wykresu z programu G*POWER z oznaczonymi poziomami błędów.*

1.2.2.2.6 Opis narzędzi pomiarowych zastosowanych moderatorów

W badaniu zdecydowano się na użycie sześciu moderatorów. Ich wybór pokierowany był chęcią otrzymania odpowiedzi na postawione w pracy pytania badawcze. Każde z narzędzi pomiarowych jest powszechnie stosowane w badaniach, głównie psychologicznych. Wykorzystano więc gotowe formularze, przetłumaczone na język polski.

Pierwsze z zastosowanych narzędzi pomiarowych to moderator powszechnie nazywany kwestionariuszem lęku uogólnionego – GAD-7[61]. Ankieta składa się z siedmiu pytań, a jej wybór pokierowany był chęcią zebrania odczuć lękowych na procesy percepcji nagrań deepfake. Generalized Anxiety Disorder (GAD-7) to narzędzie pomiarowe służące do oceny poziomu lęku u badanego[62]. Kwestionariusz składa się z pytań dotyczących różnych objawów lęku, takich jak niepokój, brak koncentracji, nadmierne zamartwianie się i inne. Badany ma za zadanie oceniać, w jakim stopniu dane objawy opisują jego osobiste doświadczenia, na skali od 0 do 3, gdzie 0 oznacza „wcale nie", a 3 oznacza „cały czas". GAD-7 jest uważane za jedno z najczęściej stosowanych narzędzi do diagnozowania lęku uogólnionego i jest szeroko wykorzystywane w różnych dziedzinach, takich jak psychiatria, psychologia czy medycyna ogólna. Wyniki tego kwestionariusza są używane do oceny stopnia nasilenia lęku u pacjenta oraz do monitorowania efektów jego leczenia.

[61] Opracowanie: Dr Robert L. Spitzer, Dr Janet B.W. Williams, Dr Kurt Kroenke oraz współpracownicy, z wykorzystaniem grantu oświatowego od firmy Pfizer Inc, https://polpharmadlaciebie.pl/materialy-dla-pacjenta/psychatria/kwestionariusz-gad-7 [dostęp: 01.01.2023].

[62] R. L. Spitzer, K. Kroenke, J. B. W. Williams, B. Löwe, „A brief measure for assessing generalized anxiety disorder. Archives of Internal Medicine", 166(10), 2006. s.1092-1097.

Kolejnym narzędziem pomiarowym został moderator depresyjności – PHQ-9[63], popularnie nazywanym testem na depresję. Ankieta zawiera 9 pytań i opracowana została przez naukowców z Columbia University pod koniec lat 90. Pierwotna wersja Robert J. Spitzera, Janet B.W. Williams oraz Kurta Kroenke składała się z 59 pytań[64], jednak ze względu na chęć ograniczenia czasu badania do niezbędnego minimum, zdecydowano się na użycie podstawowej, skróconej do 9 pytań wersji[65]. Patient Health Questionnaire-9 (PHQ-9) to narzędzie pomiarowe służące do oceny poziomu depresji u badanego. Jest to kwestionariusz składający się z pytań dotyczących różnych objawów depresji, takich jak brak energii, brak zainteresowań, trudności w koncentracji i inne. Badany ma za zadanie oceniać, w jakim stopniu dane objawy występują u niego, na skali od 0 do 3, gdzie 0 oznacza „wcale nie dokuczały", a 3 oznacza „niemal codziennie". PHQ-9 jest uważane za jedno z najczęściej stosowanych narzędzi do diagnozowania depresji. Wyniki tego kwestionariusza są używane do oceny stopnia jej nasilenia u pacjenta oraz do monitorowania efektów leczenia.

W dalszej części badano potrzebę poznawczego domknięcia (PDP), czyli to, w jakim stopniu osoba potrzebuje zdobycia innych informacji i poszerzenia perspektywy, by podjąć jakąś decyzję. Jest to jedno z najczęściej wykorzystywanych w badaniach narzędzi do pomiaru motywacji epistemicznej[66]. Wybór miernika wynika z chęci dowiedzenia się, czy wysoki poziom spowoduje to, iż badani będą chętni do podejmowania natychmiastowych decyzji, bez poszukiwania informacji w innych źródłach. Zastosowana w niniejszym badaniu skrócona wersja Skali Potrzeby Poznawczego Domknięcia[67] zawiera 15 pytań, natomiast oryginalne narzędzie liczy aż 32 pytania. Skala Potrzeby Poznawczego Domknięcia, (*Need for Cognitive Closure Scale* – NCCS) to narzędzie pomiarowe służące do oceny stopnia potrzeby osiągnięcia poznawczego domknięcia u badanej osoby, czyli uporządkowania i zrozumienia informacji oraz sytuacji, w których się znajduje. Oryginalne narzędzie utworzone zostało

[63] Opracowanie: Dr Robert L. Spitzer, Dr Janet B.W. Williams, Dr Kurt Kroenke oraz współpracownicy z wykorzystaniem grantu oświatowego od firmy Pfizer Inc, https://polpharmadlaciebie.pl/materialy-dla-pacjenta/psychatria/kwestionariusz-phq-9 [dostęp: 01.01.2023].
[64] R. L. Spitzer, K. Kroenke, J. B. W. Williams, 'Patient Health Questionnaire Study Group. Validity and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study", 1999 JAMA;282, s. 1737–1744.
[65] K. Kroenke, R. L. Spitzer, J. B. Williams, „The PHQ-9: validity of a brief depression severity measure", 2001 J Gen Intern Med. 16(9), s. 606-613.
[66] M. Kossowska, K. Hanusz, M. Trejtowicz, „Skrócona wersja Skali Potrzeby Poznawczego Domknięcia. Dobór pozycji i walidacja skali", Psychologia Społeczna 2012 tom 7 1 (20), s. 89-99.
[67] Tamże.

w oparciu o teorię Webseta i Krugalskiego[68] przez Kossakowską[69]. Skala ta składa się z pytań dotyczących preferencji badanego w zakresie poszukiwania informacji i podejmowania decyzji. Badany ma za zadanie oceniać, w jakim stopniu dane pytania opisują jego osobiste preferencje, na skali od 1 do 7, gdzie 1 oznacza „całkowicie się nie zgadzam", a 7 oznacza „całkowicie się zgadzam". Wyniki są używane do badania różnic w potrzebie poznawczego domknięcia pomiędzy jednostkami oraz do oceny wpływu tej potrzeby na różne aspekty zachowania i myślenia. Badanie to umożliwia zrozumienie tego, jak ludzie radzą sobie z niepewnością i brakiem jasności w różnych sytuacjach oraz jak ich potrzeby poznawcze wpływają na decyzje i poglądy.

Następnym zastosowanym moderatorem był kwestionariusz do pomiaru kodów moralnych (*Moral Foundations Questionnaire* – MFQ)[70]. Jest to stosunkowo nowe narzędzie, gdyż zaproponowane zostało po raz pierwszy przez Jesse Grahama i Jonathana Haidta[71] w 2011. Jego polska adaptacja – MFQ-pl – składa się z 30 pytań. Moral Foundations Questionnaire (MFQ) to narzędzie pomiarowe służące do badania struktury moralnej u ludzi. Zostało opracowane w celu lepszego zrozumienia tego, co kieruje ludzkim postępowaniem moralnym. MFQ składa się z pytań dotyczących różnych aspektów moralności, takich jak lojalność, sprawiedliwość czy szacunek dla autorytetów. Badany ma za zadanie oceniać, w jakim stopniu dane pytania opisują jego osobiste przekonania moralne, na skali od 1 do 7, gdzie 1 oznacza „całkowicie się nie zgadzam", a 7 oznacza „całkowicie się zgadzam". Wyniki MFQ są używane do badania różnic w strukturze moralnej pomiędzy różnymi grupami ludzi, na przykład między grupami etnicznymi czy politycznymi frakcjami. Badanie to jest uważane za ważne dla lepszego zrozumienia tego, jak ludzie formują swoje poglądy moralne i jak te poglądy wpływają na ich zachowanie.

Kolejnym wybranym moderatorem było narzędzie badające impulsywność – Barratt Impulsivity Scale (BIS), a dokładnie jego skrócona wersja – Barratt Impulsiveness Scale-Brief (BIS-Brief). Oryginalne, pełne narzędzie zostało utworzone i opublikowane przez Ernesta S. Barratta w 1959 roku i składało się z 30 pytań. Obecnie

[68] D.M. Webster, A. W. Kruglanski, „Individual differences in need for cognitive closure", Journal of Personality and Social Psychology, 67, 1994, s. 1049–1062.
[69] M. Kossowska, „Różnice indywidualne w potrzebie poznawczego domknięcia", Przegląd Psychologiczny, 46, 2003, s. 355–375.
[70] „(10) (PDF) MFQ-PL – Kwestionariusz do pomiaru kodów moralnych", https://www.researchgate.net/publication/281274953_MFQ-PL_-_Kwestionariusz_do_pomiaru_kodow_moralnych [dostęp: 01.08.2022].
[71] J. Graham, B. A. Nosek, J. Haidt, R. Iyer, K. Spassena, & P. H. Ditto, „Moral Foundations Questionnaire (MFQ)" [Database record]. APA PsycTests, 2011.

stosowana jest jedenasta wersja narzędzia, opublikowana w 1995 roku pod nazwą BIS-11[72]. Dla niniejszej pracy przyjęto spolszczoną, skróconą wersję narzędzia BIS-Brief. Jej rzetelność alfa Cronbacha odnotowano na poziomie 0,78[73]. Wersja ta składa się z ośmiu twierdzeń odnoszących się bezpośrednio do własnego życia respondenta. Cztery z nich badają impulsywność w działaniu jednostki, natomiast pozostałe cztery zdolność do samokontroli. Respondenci na siedmiostopniowej skali oceniali stopień zgody lub niezgody z danymi stwierdzeniami. Uzyskane za pomocą BIS-Brief wyniki mogą dostarczyć informacji na temat poziomu impulsywności danej osoby. Wyższe wyniki w kwestionariuszu BIS-Brief wskazują na większą impulsywność, którą rozumieć należy jako tendencję do podejmowania działań bez odpowiedniego zastanowienia się nad konsekwencjami, braku kontroli nad zachowaniem lub trudności w opanowaniu impulsów.

Ostatnim zastosowanym narzędziem pomiarowym była Skala Samooceny SES Morrisa Rosenberga – SES[74]. Polska adaptacja tej metody składa się z 10 pytań, a o jej rzetelności świadczy między innymi alfa Cronbacha równa przedziałowi 0,81–0,83. Skala samooceny SES Morrisa Rosenberga (*Rosenberg Self-Esteem Scale* – RSE) to narzędzie pomiarowe służące do oceny poziomu samooceny u badanego. Jest to kwestionariusz składający się z dziesięciu afirmatywnych stwierdzeń dotyczących pozytywnych aspektów samego siebie. Badany ma za zadanie oceniać, w jakim stopniu dane stwierdzenia opisują jego osobę, na skali od 1 do 4, gdzie 1 oznacza „zdecydowanie się nie zgadzam", a 4 oznacza „zdecydowanie się zgadzam". Skala samooceny SES Morrisa Rosenberga jest uważana za jedno z najczęściej stosowanych narzędzi do pomiaru samooceny i jest używana w różnych dziedzinach, takich jak psychologia, socjologia i pedagogika. Wyniki tego kwestionariusza są używane do diagnozowania różnych problemów emocjonalnych i behawioralnych oraz do oceny skuteczności programów i interwencji terapeutycznych.

[72] J. H. Patton, M. S. Stanford, E. S. Barratt, „Factor structure of the Barratt impulsiveness scale", Journal of Clinical Psychology, 51(6), 1995, s. 768-774.
[73] L. Steinberg, C. Sharp, M. S. Stanford, A. T. Tharp, „New tricks for an old measure: the development of the Barratt Impulsiveness Scale-Brief (BIS-Brief)", Psychological Assessment, 25, 2013, s. 216-226.
[74] M. Laguna, K. Lachowicz-Tabaczek, I. Dzwonkowska, „Skala Samooceny SES Morrisa Rosenberga – polska adaptacja metody", Psychologia Społeczna 2, 2007, s. 164–176.

1.2.2.2.7 Zastosowane statystyki opisowe

W trakcie opisu badania (rozdziały 4 – 6) użyto terminologii właściwej dla eksperymentu badawczego. Celem wyciągnięcia prawidłowych wniosków oraz przeanalizowania i usystematyzowania ich, większość użytego słownictwa wywodzi się ze statystyki i raportowania wyników w psychologii. Ponieważ w całej pracy struktura rozumowania i analizy została ujednolicona, poniżej opisywane są kolejno dokonywane czynności analityczne.

Statystyki opisowe, służą do opisywania danych za pomocą różnych miar i wskaźników. Mogą być wykorzystywane do prezentacji danych w sposób zrozumiały dla odbiorców, co umożliwia szybką interpretację informacji zawartych w danych. Przykładowymi miarami statystyk opisowych są średnia, mediana, modalna, odchylenie standardowe i skośność. Te miary pozwalają na opisanie charakterystyki danych, w tym ich rozkładu, skupienia lub rozproszenia. Statystyki opisowe stosowane są w wielu dziedzinach, takich jak badania marketingowe, ekonomia, medycyna czy nauki społeczne.

W niniejszej pracy statystyki opisowe prezentowane są w ujednoliconych tabelach. Kolejne wiersze tabeli odnoszą się do kolejnych statystyk opisowych, których skróty wytłumaczone zostały zarówno poniżej, jak i w załączonym do pracy opisie skrótów.

- N – liczba osób badanych, które odpowiedziały na dane pytanie.
- Brakujące odpowiedzi – liczby osób, które brały udział w badaniu, ale nie udzieliły odpowiedzi na to pytanie.
- *M* – wartość średniej – odpowiedzi na to pytanie (miara tendencji centralnej – mocno obciążona wynikami skrajnymi, czyli wysokimi i niskimi).
- *SE* – błąd standardowy średniej (mówiący o niepewności estymacji średniej w populacji, na podstawie średniej z próby badawczej). Błąd standardowy średniej to miara tego, jak bardzo średnia otrzymana w badaniu może różnić się od prawdziwej wartości średniej w populacji. Błąd standardowy średniej jest obliczany jako odchylenie standardowe całej próby podzielone przez pierwiastek liczby próbek. Im mniejszy jest błąd standardowy średniej, tym lepiej prezentują się nasze wyniki i tym większa szansa, że nasza średnia jest dobrą estymacją prawdziwej

wartości. Błąd standardowy średniej pozwala ocenić jakość naszych wyników i podejmować odpowiednie działania w celu poprawy jakości badań.

- 95% CI dolna granica przedziału ufności dla średniej i 95% CI górna granica przedziału ufności dla średniej – wskaźniki podające minimalne i maksymalne rzeczywiste wartości średniej dla populacji, z 95% prawdopodobieństwem (odpowiedzi na dane pytanie) prawdziwa średnia mieści się w przedziale między tymi dwoma wartościami. Są to wartości, które określają, jaki przedział wartości obejmuje nasza średnia z prawdopodobieństwem 95%. Innymi słowy, jeśli powtórzymy nasze badanie wielokrotnie, to w 95% przypadków średnia będzie zawarta w tym przedziale. Dolna granica przedziału ufności jest niższą wartością tego przedziału, a górna granica przedziału ufności jest wyższą wartością tego przedziału. Przedział ufności pozwala nam ocenić, jak dokładne są nasze wyniki i jak duże jest prawdopodobieństwo, że prawdziwa wartość znajduje się w tym zakresie.
- *Me* – mediana – wartość środkowa odpowiedzi na dane pytanie (połowa osób badanych odpowiadając, zaznaczyła wyższą wartość, a połowa – niższą).
- *D* – dominanta – najczęściej wybierana wartość odpowiedzi na dane pytanie.
- *SD* – odchylenie standardowe (miara zróżnicowania – mówiąca o rozrzucie wyników wokół średniej).
- *Min* – minimalna wartość odpowiedzi na dane pytanie – zaznaczona przez przynajmniej jedną osobę badaną.
- *Max* – maksymalna wartość odpowiedzi na dane pytanie – zaznaczona przez przynajmniej jedną osobę badaną.
- *SKE* – skośność (trzeci moment centralny) – miara asymetrii rozkładu wyników, informującą o tym, w jakim stopniu rozkład odchyla się od symetrii wokół wartości średniej. Wartości bliższe zera świadczą o większej symetryczności rozkładu, natomiast wartości dalsze od zera (zarówno dodatnie, jak i ujemne) świadczą o niesymetrycznym rozkładzie wyników. Wartości dodatnie mówią o prawostronnym rozkładzie

(rozkładzie prawoskośnym) – przewadze wyników niskich, natomiast wyniki ujemne mówią o lewostronnym rozkładzie (rozkładzie lewoskośnym) – przewadze wyników wysokich.

- SE_{SKE} – błąd standardowy skośności – mówi o pewności estymacji skośności z próby na populację – im większa jego wartość bezwzględna, tym z niższą ufnością traktujemy wartość skośności. Błąd standardowy skośności to wartość, która mówi nam o tym, jak bardzo rozkład danych różni się od rozkładu normalnego.
- *K* – kurtoza (standaryzowany moment średniej czwartego rzędu) – miara koncentracji wyników wokół średniej. Dodatnie wyniki świadczą o rozkładnie leptokurtycznym (większej koncentracji wyników wokół średniej); natomiast wartości ujemne świadczą o rozkładzie platokurtycznym (mniejszej koncentracji wyników wokół średniej). Innymi słowy, kurtoza mówi nam o tym, czy dane są bardziej, czy mniej spłaszczone. Jeśli dane są rozkładu normalnego, to kurtoza będzie równa 0. Kurtoza jest ważna, ponieważ wiele metod statystycznych wymaga, aby dane były zbliżone do rozkładu normalnego, aby ich wyniki byłe rzetelne.
- SE_K (Std. error K) – błąd standardowy kurtozy – mówi o pewności estymacji kurtozy z próby na populację – większa bezwzględna wartość tej miary mówi o mniejszym zaufaniu do szacowanych wartości.
- *S-W* – wartość testu Shapiro-Wilk – statystycznego testu normalności, badającego czy uzyskany rozkład danych zbliżony jest do rozkładu normalnego. Test ten polega na porównaniu rozkładu wartości z danych z teoretycznym rozkładem normalnym o tych samych wariancjach i średnich. Jeśli dane są z rozkładu normalnego, to wartość statystyki testu Shapiro-Wilk będzie bliska 1, natomiast, jeśli dane nie pochodzą z rozkładu normalnego, to wartość statystyki testu Shapiro-Wilk będzie znacząco różna od 1. Wartość statystyki testu Shapiro-Wilk jest wykorzystywana do określenia istotności statystycznej testu, czyli tego, czy dane rzeczywiście pochodzą z rozkładu normalnego.
- $p_{S\text{-}W}$ – istotność statystyczna testu Shapiro-Wilk – miara prawdopodobieństwa popełnienia błędu typu II, czyli odrzucenia hipotezy zerowej o normalnym rozkładzie danych, gdy w rzeczywistości ta

hipoteza jest prawdziwa. Wartość p testu Shapiro-Wilk jest wyrażana jako prawdopodobieństwo uzyskania wyniku testu równie ekstremalnego lub bardziej ekstremalnego od obserwowanego, zakładając prawdziwość hipotezy zerowej. Wartość p mniejsza niż poziom ufności (dla niniejszej pracy przyjęto wartość 0,05) oznacza, że istnieje wystarczająco dużo dowodów, aby odrzucić hipotezę zerową i uznać, że dane nie pochodzą z rozkładu normalnego. Należy zaznaczyć, iż odrzucenie hipotezy zerowej nie oznacza z całą pewnością, że dane nie są rozłożone normalnie. Oznacza to jedynie, że istnieje wystarczająco dużo dowodów, aby podważyć tę hipotezę.

1.2.2.2.8 Wykorzystane testy nieparametryczne

W trakcie analizy wyników badania wykonano szereg testów statystycznych, celem weryfikacji podobieństwa oraz istotności. Zastosowane w pracy testy wykorzystywane były na każdym etapie analizy zebranego materiału badawczego. Testy powtarzały się wielokrotnie i zostały wykonane kilkadziesiąt razy, celem potwierdzenia istotności statystycznej. Poniżej opisano zastosowane w pracy testy oraz uzasadniono ich wybór.

Pierwszy z zastosowanych testów to międzygrupowy test Kruskal-Wallis. Jest to nieparametryczny test statystyczny, który służy do porównywania średnich kilku grup. Test Kruskal-Wallis jest podobny do testu ANOVA, ale może być stosowany, gdy dane nie spełniają warunków koniecznych do zastosowania ANOVA, takich jak rozkład normalny lub równe wariancje w grupach. Warunkiem wykonania testu Kruskal-Wallis jest losowość danych z populacji oraz każda z badanych grup musi mieć co najmniej trzy próby niezależne (k>2)[75].

Test Kruskal-Wallis oblicza się poprzez wyliczenie średniej dla każdej grupy. Następnie należy obliczyć sumę kwadratów reszt dla każdej z grup i osobno dla całości danych. Następnie obliczamy tzw. wartość statystyki Kruskal-Wallis, która obliczana jest jako iloraz wartości H (wyliczonej na podstawie sumy rang) i iloczynu (k-1) i liczby obserwacji (N) w jednej grupie[76].

[75] D. Mider, A. Marcinkowska, „Analiza danych ilościowych dla politologów. Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP", ACAD, Warszawa 2013, s. 282.
[76] https://manuals.pqstat.pl/en:statpqpl:porown3grpl:nparpl [dostęp: 01.01.2023].

Zapis testu w formie matematycznej jest następujący[77]:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^{\,2}}{n_i} - 3(N+1)$$

H – wartość testu Kruskal-Wallis,

N – łączna liczba obserwacji we wszystkich grupach,

k – liczba porównywanych grup,

Ri – suma rang w i-tej grupie,

ni – liczba obserwacji w i-tej grupie.

Jeśli wartość statystyki Kruskal-Wallis jest duża oraz różnice między grupami są istotne statystycznie, to możemy stwierdzić, że różnice między grupami nie wynikają z przypadku. Jeśli wartość statystyki Kruskal-Wallis jest mała, oznacza to, że różnice między grupami nie są istotne statystycznie i nie możemy stwierdzić, że różnice między grupami nie wynikają z przypadku.

Test Kruskal-Wallis jest ważnym narzędziem statystycznym, ponieważ pozwala porównywać średnie kilku grup, nawet jeśli dane nie spełniają warunków koniecznych do zastosowania testu ANOVA, takich jak rozkład normalny lub równe wariancje w grupach. Jest to szczególnie ważne w badaniach, gdzie chce się ustalić, czy różnice między grupami są istotne i czy otrzymane wyniki są wiarygodne. Dzięki temu jest to bardzo przydatne narzędzie w badaniach naukowych, gdzie często trudno jest spełnić wszystkie warunki wymagane przez ANOVA. Niestety, jak zauważa D. Mider, test Kruskal-Wallis, podobnie jak inne testy nieparametryczne, cechuje mniejsza dokładność pomiaru, przez co wyniki mogą być nieco gorzej interpretowalne[78].

Kolejny wykorzystany w analizie test to test Friedmana. Jest to nieparametryczny test statystyczny, który służy do porównywania średnich kilku grup w przypadku, gdy dane są złożone i nie spełniają warunków koniecznych do zastosowania testu ANOVA z powtarzalnymi pomiarami. Test Friedmana jest często stosowany w badaniach medycznych, gdy chce się sprawdzić, czy różnice między grupami są istotne statystycznie.

[77] W. Kruskal, W. Wallis, "Use of ranks in one-criterion variance analysis", *Journal of the American Statistical Association*. 1952, **47** (260), s. 583 – 621.
[78] D. Mider, „Polacy wobec przemocy politycznej”, Warszawa, 2017, s. 201 – 202.

Aby wykonać test Friedmana, należy pobrać dane losowo z populacji. Należy pamiętać, iż dane muszą być złożone, czyli muszą składać się, z co najmniej 3 pomiarów dla każdej próbki w każdej grupie[79].

Do wykonania testu Friedmana, należy wpierw obliczyć średnią dla każdej grupy. Następnie oblicza się sumę kwadratów reszt dla każdej grupy i całości danych. Suma kwadratów reszt to suma kwadratów różnicy między każdą wartością, a średnią dla danej grupy. Następnie należy obliczyć tzw. wartość statystyki Friedmana, która obliczana jest jako iloraz 12 razy iloczynu liczby obserwacji (N) i sumy kwadratów średnich rang (R_j^2) minus iloczynu N do kwadratu, podzielone przez iloczyn liczby próbek (k) i (k-1)[80].

Podobnie do testu Kruskal-Wallis, jeśli wartość statystyki Friedmana jest duża oraz różnice między grupami są istotne statystycznie i możemy stwierdzić, że różnice między grupami nie wynikają z przypadku. Jeśli wartość statystyki Friedmana jest mała, oznacza to, że różnice między grupami nie są istotne statystycznie i nie możemy stwierdzić, że różnice między grupami nie wynikają z przypadku.

Trzecim z użytych w analizie testów jest test Durbin-Conover. Jest to statystyczny test hipotez, który służy do sprawdzenia, czy dwie grupy badanej populacji różnią się od siebie pod względem średniej. Test ten został opracowany przez Jamesa Durbina w 1951 roku i jest często stosowany w badaniach naukowych oraz w przemyśle, aby ocenić skuteczność nowych produktów lub strategii marketingowych[81].

Test Durbin-Conover jest podobny do testu t-Studenta, ale ma kilka istotnych różnic. Po pierwsze, test Durbin-Conover uwzględnia różnice w wariancjach obu grup, podczas gdy test t-Studenta zakłada, że wariancje są takie same. Po drugie, test Durbin-Conover nie wymaga rozkładu zbliżonego do rozkładu normalnego (a testy z grupy testów t-Studenta tego wymagają).

Aby przeprowadzić test Durbin-Conover, należy najpierw określić hipotezę zerową i hipotezę alternatywną. Hipoteza zerowa zakłada, że średnie obu grup są takie same, natomiast hipoteza alternatywna zakłada, że średnie są różne. Następnie należy wybrać odpowiedni poziom istotności, czyli prawdopodobieństwo błędu pierwszego rodzaju (np. 0,05 lub 0,01, dla niniejszej pracy przyjęto współczynnik 0,05).

---

[79] https://manuals.pqstat.pl/en:statpqpl:porown3grpl:nparpl [dostęp: 01.01.2023].
[80] Tamże.
[81] J. Durbin, „Incomplete blocks in ranking experiments", British Journal of Statistical Psychology, 1951, 4: 85–90

Kolejnym krokiem jest obliczenie statystyki testowej Durbin-Conovera i porównanie jej z krytycznym poziomem istotności. Jeśli statystyka testowa jest większa niż krytyczny poziom istotności, to oznacza, że istnieje istotna różnica między średnimi obu grup. W przeciwnym razie nie ma podstaw do odrzucenia hipotezy zerowej[82].

[82] PQStat - Baza Wiedzy o statystyce, testy nieparametryczne, https://manuals.pqstat.pl/en:statpqpl:porown3grpl:nparpl [dostęp: 01.01.2023].

Rozdział 2. Wpływ zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe – ustalenia terminologiczne

Chcąc przygotować pracę badawczą dotyczącą wpływu nagrań deepfake na bezpieczeństwo narodowe, nie sposób było nie sięgnąć wpierw do zgromadzonej dotychczas wiedzy dotyczącej technologii oraz mechanizmów psychologicznych wpływających na jej percepcję. Niniejszy rozdział oparty został w całości na analizie danych zastanych (zwanymi też zamiennie badaniami zza biurka lub inaczej badaniami gabinetowymi)[83]. Przybliża on dotychczasowe teorie naukowe oraz wynikające z nich wpływy na poruszany w pracy temat. W pierwszej części został on więc szczegółowo rozpisany tak, by zminimalizować rozbieżności definicyjne oraz w jak największym stopniu uszczegółowić założenia pracy.

W dalszej części zbadano literaturę dotyczącą bezpośrednio wiedzy o deepfake. Celem ujednolicenia pracy względem przyjętych w nauce o bezpieczeństwie standardów, zadano w tym miejscu niniejsze szczegółowe pytanie badawcze: jaki jest status ontologiczny wiedzy na temat wpływu zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe? Jako status ontologiczny założono sprawdzenie, czy jakakolwiek wiedza istnieje oraz jak jest ukształtowana w społeczeństwie.

Hipoteza, jaką zaproponowano, brzmi następująco: zmanipulowane materiały audiowizualne mogą mieć istotny wpływ na bezpieczeństwo narodowe, a obecny stan wiedzy na ten temat jest niepełny i nadal rozwijający się.

Weryfikacja hipotezy kontynuowana jest w trzeciej części rozdziału, gdzie przejrzana została literatura dotycząca zjawisk psychologiczno – społecznych, dotykających bezpośrednio odbiorców dezinformacji i mogących mieć znaczny wpływ na bezpieczeństwo narodowe. Sprawdzono, które elementy mogą w większym lub mniejszym stopniu wpływać na jednostkę, grupę lub całe społeczeństwo, a co za tym idzie, w jakim stopniu mogą zagrażać bezpieczeństwu narodowemu.

[83] Metodę badawczą analiza danych zastanych opisano w rozdziale pierwszym.

2.1 Sens teoretyczny tytułu

Pojęta przez Doktoranta problematyka badawcza jest niezwykle skomplikowana. Już sama nazwa obszaru badawczego, pomimo jej jednorodności, nie została jednoznacznie określona. Jako pojęcie polisemantyczne, używane jest na wiele sposobów i w wielorakim znaczeniu. Wynika to poniekąd z jego popularności i powszechnego wykorzystania we współczesnych dyscyplinach i subdyscyplinach. Część naukowców, jak na przykład Abraham Maslow, proponuje szeroką definicję bezpieczeństwa, obejmującą wiele elementów, nieujmowanych przez zwolenników etymologicznego znaczenia i odwołujących się do korzeni ewolucji.

Z tego powodu, najważniejsze pojęcia zostaną na potrzeby niniejszej dysertacji zdefiniowane w poniższym podrozdziale. Nadanie im jednolitego znaczenia wymagane jest bowiem do prawidłowego zrozumienia kolejnych rozdziałów i myśli autora. W szczególny sposób poruszone zostało znaczenie pojęcia „dezinformacja obrazem", z racji jego nietypowego zastosowania.

2.1.1 Zagrożenie

Zagrożenie rozumie się jako pośrednie lub bezpośrednie destrukcyjne oddziaływanie na dany podmiot. Stanowi on najbardziej klasyczny czynnik środowiska bezpieczeństwa, stojący obok innych trzech elementów bezpieczeństwa: szansy, wyzwania i ryzyka. W literaturze przedmiotu zagrożenia kategoryzuje się najczęściej na następujące grupy: zagrożenia potencjalne i realne; subiektywne i obiektywne; ze względu na miejsce: zewnętrzne i wewnętrzne; przedmiotowe: militarne i pozamilitarne (polityczne, ekonomiczne, społeczne, ekologiczne); charakter stosunków: kryzysowe i wojenne (lub też konfliktowe i niekonfliktowe) oraz intencjonalne i przypadkowe[84]. W opisie zagrożeń intencjonalnych wyróżnić można cztery elementy: aktora, jego intencje, możliwości oraz czas na reakcję. W tej definicji poziom zagrożenia wzrasta wraz z narastaniem wrogości przeciwnika, rozwojem jego możliwości siłowych oraz skracaniem się czasu na reakcję[85].

84 R. Jakubczak, „Obrona narodowa w tworzeniu bezpieczeństwa III RP", Dom Wydawniczy BELLONA, Warszawa 2003.
85 J. M. Fish, S. J. McCraw, Ch. J. Reddish, „Fighting in the gray zone: a strategy to close the preemption gap", US Army War College, Strategic Studies Institute, 2004, s. 4.

Stanisław Koziej uważa, że w stosunkach międzynarodowych, siłę zagrożenia możemy wyrazić, jako iloczyn potencjału materialnego i niematerialnego.

S = P x W

Gdzie S – siła stanowiąca zagrożenie, P – potencjał fizyczny przeciwnika (siły militarne, zasoby), w – jego wola działania (operacje informacyjne, działania niemilitarne, poparcie społeczne)[86]. Dezinformacja ujęta jest w woli działania i ma na celu, poprzez jej redukcję u przeciwnika, obniżyć jego siłę, a tym samym zmniejszyć zagrożenie, które stwarza.

Pojęcie zagrożenia bezpieczeństwa narodowego było dawniej definiowane w ścisłym zawiązku z przyjmowanym za priorytetowy (a nawet jedyny) militarnym obszarem bezpieczeństwa narodowego[87]. Z upływem czasu i różnymi katastrofalnymi wydarzeniami, zidentyfikowano nowe obszary bezpieczeństwa narodowego. Zbigniew Ciekanowski definiuje zagrożenie jako brak bezpieczeństwa, przez co staje się niezmienną i nieuniknioną, a w niektórych wypadkach powszechną, rzeczywistością życia ludzkiego[88]. Jednocześnie ma ono ścisły związek z bezpieczeństwem, które w ten sposób czyni zagrożenie jego podstawową kategorią. Identyfikacja zagrożeń i wiedza o nich stają się zatem podstawowym warunkiem do wszczęcia działań zapobiegawczych oraz organizacji obrony.

Ryszard Jakubczak typologizuje zagrożenia na pierwotne i wtórne. Do pierwszej grupy zalicza naturalne (katastrofy, kataklizmy), techniczne (awarie), militarne (terror) oraz nadzwyczajne zagrożenia środowiska (o charakterze antropomorficznym). Do zagrożeń wtórnych wpisują się zaś zagrożenia egzystencji człowieka (głód, pandemie), społeczne (patologie społeczne), naruszanie równowagi biologicznej (epizootie, epifitozy) oraz masowe straty (klęski żywiołowe)[89].

Michał Brzeziński zagrożenia dzieli na trzy kategorie: codzienne, sytuacje kryzysowe oraz zagrożenia nadzwyczajne. Bezpieczeństwo przy tej typologii nie oznacza

86 S. Koziej, „Bezpieczeństwo: istota, podstawowe kategorie i historyczna ewolucja", [w:] Bezpieczeństwo narodowe 18 (2), 2011, s. 25.
87 Raport o stanie systemu przeciwdziałania, zwalczania i usuwania skutków nadzwyczajnych zagrożeń dla ludzi i środowiska, Warszawa 1997, str. 28.
88 Z. Ciekanowski, „Rodzaje i źródła zagrożeń bezpieczeństwa", [w:] „Bezpieczeństwo i Technika Pożarnicza", nr. 1, Warszawa 2010, s. 29 – 31.
89 R. Jakubczak, „Obrona narodowa w tworzeniu bezpieczeństwa III RP", Dom Wydawniczy BELLONA, Warszawa 2003.

stanu braku zagrożeń, lecz jako minimalny, akceptowalny społecznie ich poziom[90]. Zagrożenia codzienne to te, którym jesteśmy w stanie samodzielnie zapobiegać posiadanymi środkami, bez mobilizacji dodatkowych sił. Zagrożenia nadzwyczajne wymagają zaś interwencji zewnętrznych podmiotów, celem ich zwalczenia. Działanie to nie jest również wynikiem wyboru, a koniecznością jego podjęcia.

Zagrożeniem nazwane zostały więc wszelkie okoliczności lub zdarzenia, które mogą działać przeciwko podmiotowi lub zasobowi w sposób, który może spowodować jego szkodę.

2.1.2 Manipulacja

Pomimo tego, że w tytule uniknięto tego sformułowania, konieczne zdaje się jego uprzednie zdefiniowanie. Pojęcie to wiele razy powtarza się w niniejszej pracy i ma dla niej głęboki sens. Ze względu na ilość różnych definicji *manipulacji w* naukach o bezpieczeństwie, konieczne jest ustalenie jednego, właściwego dla niniejszej pracy znaczenia.

„Vademecum Bezpieczeństwa"[91] definiuje dwa pojęcia, zawierające w sobie słowo „manipulacja". Jest to manipulacja informacją oraz manipulacja medialna. Pierwsze z nich, autorka tłumaczy jako „rodzaj modyfikacji dokonywanej na informacji lub procesie informacyjnym (np. Na transmisji, udostępnianiu, interpretacji i wykorzystywaniu informacji)"[92]. Drugie z pojęć – manipulacja medialna – autor definiuje jako celowe formułowanie przekazu medialnego, w taki sposób, aby jego treść służyła osiągnięciu celu założonego przez nadawcę, niezależnie od zgodności przekazu z faktami[93]. Obie definicje zawierają zbliżone do siebie elementy. Są nimi modyfikacja informacji lub jej przekazu, celowość działania oraz chęć osiągnięcia jakiejś korzyści.

Wojciech Warecki, w swojej definicji manipulacji dodaje jeszcze jeden istotny aspekt. Jest nim zamierzone wywieranie wpływu w taki sposób, by adresat manipulacji nie zdawał sobie z tego sprawy[94].

---

[90] M. Brzeziński, „O zagrożeniach codziennych, nadzwyczajnych i sytuacjach kryzysowych z perspektywy systemowej", [w:] S. Sulowski, M. Brzeziński „Trzy wymiary współczesnego bezpieczeństwa" 2015.
[91] O. Wasiuta, R. Klepka, R. Kopeć (red.), „Vademecum bezpieczeństwa", Libron, Kraków 2018.
[92] P. Motylińska „Manipulacja informacją" [w:] O. Wasiuta, R. Klepka, R. Kopeć (red.), „Vademecum bezpieczeństwa", Libron, Kraków 2018, s. 423 – 428.
[93] R. Klepka, „Manipulacja medialna" [w:] O. Wasiuta, R. Klepka, R. Kopeć (red.), „Vademecum bezpieczeństwa", Libron, Kraków 2018, s. 428 – 435
[94] M. Warecki, W. Warecki, „Słowo o manipulacji, czyli krótki podręcznik samoobrony", Poltext, Warszawa 2006.

Stąd zaproponować można zwartą definicję manipulacji jako działania lub zachowania, które ma na celu wpływanie na myślenie, uczucia lub zachowanie innej osoby, zwykle w celu osiągnięcia korzyści dla samego manipulującego, bez świadomości bycia celem manipulacji. Działanie to może przybierać różne formy, takie jak oszukiwanie, kłamstwo, przekonywanie, perswadowanie, nacisk, groźby lub szantaż.

Manipulacja jest często stosowana w relacjach międzyludzkich, zarówno w życiu prywatnym, jak i zawodowym. Może być używana przez osoby, które chcą wpływać na decyzje innych lub osiągać swoje cele kosztem innych. Osoba manipulująca może używać różnych taktyk, takich jak podawanie fałszywych informacji, wykorzystywanie emocji innych ludzi czy stosowanie szantażu emocjonalnego.

W przypadku tworzenia filmów deepfake manipulacja przejawiać się będzie poprzez tworzenie fałszywego przekazu, celem oszukania odbiorcy i wpłynięcia na jego myślenie, uczucia lub wprost konkretne zachowania. Te trzy elementy wskazują największe wyzwania związane z manipulacjami wideo oraz ich wpływem na bezpieczeństwo narodowe.

Odpowiednio przygotowane nagranie, obarczone odpowiednim kontekstem, może prowadzić do negatywnych skutków dla obu stron. Osoba manipulowana może poczuć się zmanipulowana i oszukana, czego konsekwencją są poczucie winy, wstydu lub brak zaufania do innych. Z drugiej strony, osoba manipulująca może tracić szacunek i zaufanie innych, a także narażać się na konsekwencje prawne, jeśli jej działania są nielegalne.

2.1.3 Dezinformacja multimedialna z wykorzystaniem technologii deepfake

Dezinformację rozpatruje się na ogół w jej wąskim znaczeniu jako sfabrykowane świadectwo, wprowadzające w błąd. Powstało ponadto specyficzne słownictwo wykorzystywane na określenie zjawisk w Internecie: boty (*bots*), trolle (*trolls), fake news, cyborgs, flame wars, echo chambers, filter bubbles, deepfake, sockpuppet, straw man, atak sybill, astroturfing*, etc.

Tomasz Kacała definiuje dezinformację jako celowe przekazywanie nieprawdziwych informacji. Ma to na celu osiągnięcie określonego efektu – wpłynięcia na odbiorcę[95]. W rozważaniach nad dezinformacją wyłaniają się dwa dominujące nurty.

[95] T. Kacała, „Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa", Przegląd Prawa Konstytucyjnego, nr 2, 2015, s. 49–66.

Pierwszy z nich pojmuje pojęcie w węższym znaczeniu jako celowe wprowadzenie w błąd grupy społecznej lub większej populacji, przy wykorzystaniu szerokiego wachlarza narzędzi i technik. W szerszym znaczeniu dezinformacją nazywane są techniki wpływania na ludzi, celem wywołania określonych zmian w zachowaniu, poglądach społecznych[96].

Na potrzeby niniejszej pracy dezinformacją nazwane zostały procesy celowego wprowadzania w błąd podmiotu, przy pomocy fałszywych lub zmanipulowanych informacji w celu manipulowania opinią publiczną lub osiągnięcia własnych korzyści (na przykład: politycznych, ekonomicznych, społecznych czy militarnych).

Dezinformacja jest więc specjalnym rodzajem manipulacji, w której wprowadza się ludzi w błąd poprzez rozpowszechnianie fałszywych lub wybiórczych informacji. Manipulacja jest bardziej ogólnym pojęciem, obejmującym wprowadzanie w błąd poprzez różnego rodzaju działania, w tym nie tylko poprzez rozpowszechnianie dezinformacji, ale też przez nieuczciwe argumenty i strategie wpływu.

Warto podkreślić, iż dezinformacja jako zjawisko, od kilku lat podnoszone jest na wokandzie jako jedno z głównych zagrożeń nadchodzących czasów[97]. Autorzy raportu „The Global Risks Report 2023" jednoznacznie obwiniają dezinformację jako przyczynę spadku ilości szczepień, między innymi na covid-19 czy polio, a co za tym idzie wzrost śmiertelności. Autorzy publikacji wskazują, iż poprzez dezinformację łatwo można pogłębiać społeczne podziały, doprowadzić do erozji spójności społecznej oraz kryzysów finansowych. Sumarycznie przekłada się to na spadek bezpieczeństwa narodowego oraz zubożenie społeczne.

Jedną z technik dezinformacji jest dezinformacja multimedialna. Poprzez pojęcie multimedialna – użyte w tytule dysertacji – rozumie się nie tylko obraz, lecz także dźwięk. Ten zabieg wynika z braku w języku polskim odpowiednika słowa *deepfake*[98], które nawet w języku angielskim, z racji swojego młodego stażu, pozostaje cały czas niedookreślone. Naukowcy badający to zagadnienie najczęściej poprzez *deepfake* określają technologię głębokiego uczenia maszynowego (*deep learning*) do tworzenia lub

[96] V. Volkoff, Dezinformacja – oręż wojny, Warszawa 1991, s. 8. Cyt. Za: T. Kacała, „Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa", Przegląd Prawa Konstytucyjnego, nr 2, 2015, s. 49–66.
[97] Raport „The Global Risks Report 2023", 18th Edition, World Economic Forum.
[98] Pojęcie *deepfake* oraz opis technologii rozwinięte zostało w kolejnych podrozdziałach.

manipulowania treściami wideo – obrazem i dźwiękiem[99]. Celem takiego działania jest ukazanie sytuacji, która nigdy nie miała miejsca, umieszczenie twarzy czy głosu danej osoby w nagraniu, w którym nigdy nie wystąpiła.

Dezinformacją multimedialną nazywa się więc wykorzystanie obrazów i dźwięków przez podmioty prowadzące dezinformację do celowego przedstawiania wprowadzającego w błąd lub sfabrykowanego, nieprawdziwego obrazu rzeczywistości.

Ze względu na sposób działania, wyróżniona została poniższa typologia działania *deepfake w* zakresie dezinformacji obrazem:

1. Manipulowanie kontekstem np. poprzez łączenie rzeczywistych obrazów z wprowadzającymi w błąd podpisami
2. Kadrowanie obrazów, celem zwiększenia wyrazistości wybranych fragmentów lub ukrycia pewnych fragmentów
3. Manipulowanie wizualizacją w celu przedstawienia innej rzeczywistości
4. Wytwarzanie treści poprzez łączenie zmanipulowanych obrazów ze zmanipulowanym tekstem
5. Ingerencja w obraz, dodanie lub usunięcie wybranych fragmentów

Pierwsze nagrania wykorzystujące technologię *deepfake* utworzone zostały przez użytkownika Reddit[100], o tym samym pseudonimie i zaprezentowane przez niego na tym portalu w październiku 2017 roku. Następnie kod aplikacji został udostępniony szerszej grupie użytkowników na GitHubie, przez co internetowa społeczność oddolnie rozwinęła inicjatywę, tworząc kilka aplikacji, automatyzujących i upraszczających ten proces.

Pierwsze zmanipulowane nagrania, trudne do rozpoznania, pojawiły się jednak dopiero w 2020 roku. Do tego czasu szacuje się, że 90% powstałych materiałów to filmy pornograficzne. Dopiero od 2020 roku rozpoczęto manipulowanie na większą skalę wypowiedzi polityków, celebrytów czy innych liderów wpływu społecznego. Narzędzie początkowo stworzone do zabawy, stało się obecnie narzędziem dezinformacyjnym,

[99] I. Dąbrowska, „Deepfake – nowy wymiar internetowej manipulacji", Zarządzanie Mediami. 8. 2020, s. 89-101.

[100] Reddit – serwis internetowy przedstawiający linki do różnorodnych informacji, które ukazały się w Internecie. Serwis jest głównie anglojęzyczny, chociaż jego interfejs został przetłumaczony na wiele języków. Reddit ma około 180 milionów zarejestrowanych użytkowników i zajmuje 19. pozycję w rankingu Alexa Internet. Link do portalu: https://www.reddit.com/, [dosęp: 01.01.2023].

mogącym w najbliższej przyszłości doprowadzić do zachwiania się obecnego modelu zaufania do mediów, zwłaszcza internetowych[101].

Przewiduje się tworzenie zinstytucjonalizowanego i wysoce sprofesjonalizowanego przemysłu dezinformacyjnego przekraczającego granice państw i skutkującego realnymi negatywnymi, bliższymi i dalszymi konsekwencjami w zakresie polityki[102]. Przemysł ten objąłby swoim działaniem zwłaszcza obszar mediów audiowizualnych. Stanowiłby on zagrożenie zwłaszcza dla państw demokratycznych, najbardziej podatnych na dezinformację.

Duże znaczenie zdają się mieć ze względu na swój zasięg oraz syntetyczne wnioski, wszelkiego rodzaju metaanalizy obejmujące próby analiz odnoszących się do globalnego przemysłu dezinformacyjnego. W kategoriach porównawczych zagadnienie to rozpatrują Henry Schneier i Bruce Farrell analizując dwa typy idealne systemów politycznych – demokracje (*democracy*) i autokracje (*autocracies*), jako systemy informacyjne[103]. Według autorów demokracje są systemami o wiele bardziej podatnymi na dezinformację, przekazy zalewające debatę publiczną i zakłócające wspólne procesy polityczne. Przekonanie o szczególnej podatności demokracji na walkę i wojnę informacyjną z użyciem Internetu podzielają Herbet Lin i Jackie Kerr, wskazując, że demokracje, zarówno stare, jak i nowe są podatne na wszystkie trzy elementy repertuaru wojny informacyjnej: operacji propagandowych (*propaganda operations*), operacji wyciekowych (*leak operations*) oraz operacji wywołujących chaos (*chaos-producing operations*)[104]. Studium porównawcze systemów medialnych Rosji i Stanów Zjednoczonych prowadzi Sara Oates do tożsamych konkluzji: system libertariański (*the libertarian system*) przerzucający na obywateli ciężar odróżniania faktów od fikcji jest istotnie bardziej podatny, niż system autorytarny/neoradziecki (*authoritarian system*), działający na rzecz i pod kontrolą instytucji państwowych[105].

[101] O. Wasiuta, S. Wasiuta, „Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość", Studia de Securitate 9(3). 2019.
[102] D. Pogue, „How to stamp out fake news", Scientific American, 20316(2), 2017, s. 24.
[103] H. Farrell, B. Schneier, „Common Knowledge Attacks on Democracy" 2018.
[104] H. Lin, J. Kerr, „On Cyber-Enabled Information/Influence Warfare and Manipulation", Oxford University Press: 2018 forthcoming, 2018.
[105] S. Oates, „When Media Worlds Collide: Using Media Model Theory to Understand How Russia Spreads Disinformation in the United States", 2018.

2.1.4 Bezpieczeństwo

Bezpieczeństwo jako pojęcie w naukach o bezpieczeństwie ujmuje się, jako stan "bez pieczy". Polskie słowo wywodzi się od łacińskiego *securitas*, czyli *se* – oddzielnie, osobno lub *sine* – bez oraz *cura* – troska, opieka, dbanie o coś lub kogoś[106]. W języku angielskim odpowiednikiem jest *security,* jednak czasami używa się pojęcia *safety,* różniącego się jednak w znaczeniu. *Security* polega na tworzeniu ochrony przed ryzykiem lub niebezpieczeństwem, podczas gdy *safety* oznacza stan bycia wolnym przed niebezpieczeństwem lub zagrożeniami[107]. Pojęcie na język polski można więc tłumaczyć jako „wolny od trosk, bezpieczny". Definiuje się to pojęcie również jako wolność od strachu i zagrożeń, takich jak atak. Przeciwnymi pojęciami są niebezpieczeństwo i stan zagrożenia.

Szkoła kopenhaska, proponująca rozszerzoną koncepcję bezpieczeństwa, kategoryzuje pojęcie na pięć elementów: bezpieczeństwo militarne, polityczne, ekonomiczne, społeczne i ekologiczne[108]. Waldemar Kitler zauważa jednak, iż powyższy podział szczegółowy nauk o bezpieczeństwie z 1998 roku należy uzupełnić o kolejne grupy. W. Kitler proponuje podział na nauki o obronności, bezpieczeństwie publicznym, bezpieczeństwie powszechnym, bezpieczeństwie politycznym, bezpieczeństwie ekonomicznym, ekologicznym oraz społecznym[109]. Ponadto wymienia nauki o bezpieczeństwie informacyjnym oraz o bezpieczeństwie w cyberprzestrzeni. W. Kitler zwraca również uwagę, iż nie jest to podział zamknięty, wyróżniając kategorię „inne". W związku z mnogością podziałów bezpieczeństwa, w niniejszej pracy zdecydowano się zastosować dychotomiczny podział bezpieczeństwa na bezpieczeństwo personalne (rozdział 5) i strukturalne (rozdział 6)[110].

Bezpieczeństwo personalne powszechnie rozumie się jako stan wolności jednostki od zagrożeń, takich jak przemoc, przestępczość, wypadki, choroby czy katastrofy naturalne. Obejmuje ono ochronę życia, zdrowia, dobrostanu i praw jednostki.

106 B. Wiśniewski (red.), „Od nauk wojskowych do nauk o bezpieczeństwie", Szczytno 2014, s. 6-9.
107 O. V. Bubnovskaia, V. V. Leonidova, A. V. Lysova, „Security or Safety: Quantitative and Comparative Analysis of Usage in Research Works Published in 2004–2019", Behavioral Sciences 2019.
108 B. Buzan, O. Wæver, J. de Wilde, „Security: A New Framework for Analysis", CO: Lynne Rienner, Londyn 1998.
109 Kitler, Waldemar. „Nauki o bezpieczeństwie w systemie dziedzin i dyscyplin naukowych w Polsce", [w:] „Nauki o bezpieczeństwie: poszukiwanie podstaw", Akademia Sztuki Wojennej, Warszawa, 2022, s. 105-121.
110 K. Drabik, „Bezpieczeństwo personalne i strukturalne", Warszawa 2013.

Szeroko rozumiane bezpieczeństwo personalne zakłada stworzenie człowiekowi takich warunków egzystencji, które zapewnią mu swobodny rozwój[111]. W niniejszej pracy poprzez bezpieczeństwo personalne rozumieć należy ochronę jednostki przed krzywdą, taką jak przestępstwa, przemoc, manipulacja czy oszustwa.

Mówiąc o bezpieczeństwie strukturalnym, najczęściej rozumie się je jako ochronę systemów i instytucji przed zagrożeniami, takimi jak katastrofy naturalne, awarie techniczne, ataki terrorystyczne lub konflikty zbrojne. Bezpieczeństwo strukturalne obejmuje działania mające na celu zapobieganie tym zagrożeniom, łagodzenie ich skutków i szybkie reagowanie na nie. Janusz Świniarski zwraca szczególną uwagę na stronę organizacyjną i instytucjonalną życia społecznego w kontekście międzynarodowym, regionalnym, państwowym i lokalnym. Zdaniem autora, istotą bezpieczeństwa strukturalnego jest takie ukierunkowanie działalności wszystkich instytucji życia społecznego, właściwych dla wielorakich jego wymiarów, aby ich działanie, a przede wszystkim jego efekty, gwarantowały bezpieczeństwo personalne[112]. W niniejszej pracy pod pojęciem bezpieczeństwa strukturalnego w rozdziale szóstym zaproponowane zostały modele i strategie przeciwdziałania dezinformacji. Zgodnie z definicją mają one na celu podniesienie odporności społeczeństwa na dezinformację, a przez to wzmocnienie bezpieczeństwa personalnego.

Bezpieczeństwo rozumiane jest często jako stan. Problemem, jaki generuje takie podejście, jest niemierzalność – różne jest jego postrzeganie zarówno przez społeczeństwo, jak i władze państwa. Ponadto postrzeganie stanu bezpieczeństwa jest bardzo subiektywne – każda jednostka odczuwa je w odmienny sposób. Ciemność towarzysząca ludziom przez jedną trzecią doby jest naturalna. Pomimo tego, dla wielu osób – nyktofobów – powoduje odczuwanie lęku i utratę poczucia bezpieczeństwa. Daniel Frei wyróżnia następujące cztery stany bezpieczeństwa: stan braku bezpieczeństwa – występuje realne zagrożenie, a jego postrzeganie jest prawidłowe; stan obsesji – nieznaczne zagrożenie postrzegane jest jako duże; stan fałszywego bezpieczeństwa – zagrożenie jest poważne, jednak marginalizuje się jego znaczenie; stan

[111] M. Sokołowski, „O pojęciu i istocie bezpieczeństwa personalnego", Kultura Bezpieczeństwa Nr 33, 2019, s. 117–130.
[112] J. Świniarski, „O naturze bezpieczeństwa", Warszawa – Pruszków 1977, s. 13.

bezpieczeństwa – najbardziej pożądany – zagrożenie jest niewielkie, a jego postrzeganie prawidłowe[113].

Odmienne od powyższego prezentują zwolennicy ujęcia procesualnego. Bezpieczeństwo rozumiane jest jako proces – ciągła działalność ludzi, społeczeństwa w tworzeniu stanu bezpieczeństwa. Arkadiusz Sekściński zauważa, że w naukach społecznych ciężko jest mówić o bezpieczeństwie, jako o czymś stałym i niezmiennym[114].

Bezpieczeństwo ujęte jest również jako jedna z najważniejszych potrzeb egzystencjalnych, w piramidzie potrzeb Maslowa stanowi element składowy drugiego poziomu, zaraz po potrzebach fizjologicznych[115]. Potrzeba ta wynika z obiektywnych warunków bytowania ludzi i różnych grup społecznych oraz ich wzajemnych relacji, wymagając troski o jej zaspokojenie. W tym rozumieniu chodzi nie tylko o przetrwanie, lecz również bezpieczeństwo rozwoju czy wzbogacanie jednostki.

Ryszard Zięba rozróżnia pojęcie bezpieczeństwa w znaczeniu wąskim i szerokim. Wąskie rozumienie to brak zagrożeń, szerokie zaś to aktywne kształtowanie wewnętrznej i zewnętrznej polityki państwa, tak by zapewnić jego przetrwanie i możliwość rozwoju[116]. R. Zięba zaznacza również, iż bezpieczeństwo jest podstawową potrzebą grup społecznych oraz jednostek wchodzących w ich skład. Co za tym idzie, jest również podstawową potrzebą państwa i systemów międzynarodowych[117].

Zgoła odmienne podejście, od głównego nurtu rozważań prezentuje Leszek Korzeniowski. Definiuje on bezpieczeństwo jako zdolność do kreatywnej aktywności jednostki – możliwości posiadania, rozwoju i funkcjonowania[118].

Podsumowując powyższe rozważania, zauważyć należy, iż w toku pogłębionych rozważań naukowców poszerzony został zakres podmiotowy (zbiór wartości chronionych) i przedmiotowy bezpieczeństwa (odejście od spojrzenia na bezpieczeństwo

[113] D. Frei, „Grundfragen der Weltpolitik", Stuttgart 1977. Cyt. za: M. Adamczyk, „Teoretyczne wprowadzenie do badań nad bezpieczeństwem", [w:] M. Debita, M. Adamczyk (red.), „Polska – Europa – Świat. Wczoraj i dziś", Poznań 2017, s. 59 – 62.
[114] A. Sekściński, „Bezpieczeństwo wewnętrzne w ujęciu teoretycznym. Geneza i współczesne rozumienie w naukach politycznych", e-Politikon 2013, nr 6, s. 42.
[115] A. Maslow, „Motywacja i osobowość", Warszawa 1990, s. 140 – 149.
[116] R. Zięba, „Pojmowanie bezpieczeństwa międzynarodowego w XXI wieku", [w:] R. Zięba (red.), „Bezpieczeństwo międzynarodowe w XXI wieku", s. 17 – 24.
[117] R. Zięba, „Pojęcie i istota bezpieczeństwa państwa w stosunkach międzynarodowych", „Sprawy Międzynarodowe" nr. 10, 1989, s. 50.
[118] L. F. Korzeniowski, „Podstawy nauk o bezpieczeństwie", Warszawa 2012, s. 71 – 79.

przez pryzmat państwa na rzecz jednostki). Duże znaczenie przykłada się również do subiektywnego charakteru bezpieczeństwa oraz uwspólnotowienia tego pojęcia.

2.1.5 Bezpieczeństwo narodowe

W definiowaniu pojęć ujęcie indywidualne bezpieczeństwa i narodu prowadzić może do niezrozumienia związku frazeologicznego, który budują. W odczuciu autora należy więc po uprzednim zdefiniowaniu bezpieczeństwa, określić również znaczenie zwrotu bezpieczeństwo narodowe. Jest to bowiem pojęcie funkcjonujące w naukach o bezpieczeństwie: jest ono przedmiotem wielu dyskusji i dysertacji, a jego znaczenie tematem rozważań naukowych. Prawidłowe zdefiniowanie pojęcia pozwoli wyznaczyć obszary, na jakie dezinformacja ma współcześnie oddziaływać oraz pozwoli w prawidłowy sposób wyznaczyć obszar badań.

Bezpieczeństwo narodowe (*national security*) najczęściej rozumiane jest, jako jedna z podstawowych funkcji każdego państwa, która obejmuje problematykę przeciwstawienia się wszelkim zagrożeniom zewnętrznym oraz wewnętrznym dla istnienia oraz rozwoju narodu i państwa. Państwo w trosce o własne bezpieczeństwo narodowe ustala zbiór wartości wewnętrznych, które jego zdaniem powinny być chronione przed zagrożeniami. Bezpieczeństwo narodowe postrzegane jest zatem jako zdolność narodu (państwa) do obrony terytorium i wartości.

W naukach o bezpieczeństwie przyjmuje się, że pojęcia bezpieczeństwo narodowe i bezpieczeństwo międzynarodowe mają charakter umowny[119]. Andrzej Wawrzusiszyn uważa, że bezpieczeństwo narodowe stanowi obecnie pierwotną, egzystencjalną oraz naczelną potrzebę i wartość jednostek, społeczeństw oraz państwa. Jest to priorytetowy cel wszelkich działań organów państwa. Państwo troszczy się o bezpieczeństwo narodowe, poprzez realizację dwóch kluczowych celów – obronę i ochronę wartości, interesów narodowych oraz tworzenie wewnętrznych i zewnętrznych okoliczności swobodnego rozwoju[120].

Wojciech Kotowicz wyprowadza pojęcie bezpieczeństwa narodowego z egzystencjalnych potrzeb i interesów ludzkich, społeczności zorganizowanej w państwo. W jego opinii stan bezpieczeństwa narodowego to stan umożliwiający

[119] A. Wawrzusiszyn (red.), „Praca dyplomowa z bezpieczeństwa – wprowadzenie do badań", Warszawa 2016, s. 70.
[120] W. Pokruszyński, „Teoretyczne aspekty bezpieczeństwa", Józefów 2010, s. 10 – 11, Cyt. Za: A. Wawrzusiszyn (red.), „Praca dyplomowa z bezpieczeństwa – wprowadzenie do badań", Warszawa 2016, s. 71.

niezakłóconą egzystencję i rozwój państwa. W. Kotowicz przytacza definicję Stanisława Lipskiego, mówiącą, iż bezpieczeństwo narodowe to zdolność przeciwstawienia się państwa wszystkim potencjalnym lub istniejącym zagrożeniom[121].

W. Kotowicz dokonuje również kategoryzacji elementów składających się na bezpieczeństwo narodowe. Wskazuje, iż dany rodzaj bezpieczeństwa składowego, odpowiada podobnemu rodzajowi zagrożeń. W. Kotowicz wyróżnia podległe: bezpieczeństwo polityczne, energetyczne, społeczne, militarne i ekonomiczne. Składowa ocena wymienionych elementów tworzy ocenę zagrożenia państwa[122].

Pojęciem pokrewnym do bezpieczeństwa narodowego, jest bezpieczeństwo państwa. Ryszard Zięba zaznacza, iż pokrewieństwo to jest bliskie do tego stopnia, iż w naukach zachodnich pojęcia te nie znajdują rozróżnienia i traktowane są identycznie[123]. W najszerszym rozumieniu obejmuje ono wymiary wewnętrzne, jak i zewnętrzne, dotykając nie tylko polityki państwa, lecz także jego stosunków międzynarodowych[124]. Definicja bezpieczeństwa państwa ujęta została w konstytucji i w przeciwieństwie do bezpieczeństwa narodowego posiada klarowną definicję. W odczuciu autora bezpieczeństwo narodowe jest jednak szersze w swoim zakresie, obejmując inne rodzaje bezpieczeństwa, nieujęte przez bezpieczeństwo państwa. Jest to bezpieczeństwo obywateli oraz cała typologia bezpieczeństwa wewnętrznego i zewnętrznego[125]. Na gruncie krajowym można więc przyjąć, że bezpieczeństwo narodowe jest pojęciem o najszerszym zakresie przedmiotowym, które obejmuje również inne rodzaje bezpieczeństwa, jak właśnie bezpieczeństwo obywateli, bezpieczeństwo wewnętrzne oraz bezpieczeństwo zewnętrzne. Poprzez to rozumie się zdolność państwa do przeciwstawienia się wszelkim zagrożeniom zewnętrznym jak i wewnętrznym, dla

---

121 S. Lipski, „Bezpieczeństwo narodowe – wybrane zagadnienia terminologiczne", [w:] T. Jemioło, K. Rajchel, „Bezpieczeństwo narodowe i zarządzanie kryzysowe w Polsce w XXI wieku – wyzwania i dylematy: praca zbiorowa", Warszawa 2008, s. 39, Cyt. Za: W. Kotowicz, „Bezpieczeństwo narodowe", [w:] A. Żukowski (red.), M. Hartliński (red.), W. T. Modzelewski (red.), J. Więcławski (red.), „Podstawowe kategorie bezpieczeństwa narodowego", Olsztyn 2015, s. 133.

122 W. Kotowicz, „Bezpieczeństwo narodowe", [w:] A. Żukowski (red.), M. Hartliński (red.), W. T. Modzelewski (red.), J. Więcławski (red.), „Podstawowe kategorie bezpieczeństwa narodowego", Olsztyn 2015, s. 134 – 138.

123 R. Zięba, „Pojęcie i istota bezpieczeństwa państwa w stosunkach międzynarodowych", „Sprawy Międzynarodowe" nr. 10, 1989, s. 8.

124 R. Wróblewski, „Podstawowe pojęcia z dziedziny polityki bezpieczeństwa, strategii i sztuki wojennej", Warszawa 1993, s. 9 – 10.

125 M. Nowiński, „Pojęcie bezpieczeństwo narodowe w prawie europejskim i międzynarodowym w kontekście uprawnień służb specjalnych", [w:] „Uprawnienia Służb Specjalnych Z Perspektywy Współczesnych Zagrożeń Bezpieczeństwa Narodowego. Wybrane Zagadnienia", Warszawa 2017.

istnienia i rozwoju narodu oraz państwa. Państwo dla ochrony bezpieczeństwa narodowego tworzy zbiory wartości wewnętrznych, a następnie obejmuje je swoją troską. Definiować bezpieczeństwo narodowe należy więc jako sumę bezpieczeństwa państwa i bezpieczeństwa wewnętrznego. W tym kontekście pierwsze pojęcie stanowi o obronie terytorium, zaś drugie wartości ważnych dla narodu, również w znaczeniu militarnym[126].

Brak precyzyjnego i jasnego określenia znaczenia pojęcia bezpieczeństwa narodowego wiąże się także z problematycznym rozdziałem kompetencyjnym w obszarze ochrony bezpieczeństwa i porządku publicznego, czyli – innymi słowy – kompetencjami o charakterze policyjnym oraz w obszarze dedykowanym służbom specjalnym, czyli prowadzeniu czynności o charakterze wywiadowczym (oraz kontrwywiadowczym). W niektórych państwach członkowskich Unii Europejskiej omawiane kompetencje nie są rozdzielone i mają charakter „mieszany", czyli np. służba specjalna poza kompetencjami wywiadowczymi i kontrwywiadowczymi ma również uprawnienia o charakterze policyjnym.

Warto wskazać, że ani w przepisach prawa Unii Europejskiej, ani w dotychczasowym orzecznictwie TSUE dotychczas nie wypracowano spójnej, precyzyjnej i jednoznacznej definicji pojęcia bezpieczeństwo narodowe. Ponadto należy pamiętać, że zarówno na forum Unii Europejskiej, jak i państw członkowskich pojawiają się również inne pojęcia, znaczeniowo zbliżone do "bezpieczeństwa narodowego", które również nie są precyzyjne zdefiniowane. Są to pojęcia zasadniczo powiązane z szeroko rozumianym bezpieczeństwem: bezpieczeństwo wewnętrzne, bezpieczeństwo państwa, bezpieczeństwo publiczne i bezpieczeństwo obronne. Z uwagi na to, że wszystkie wymienione terminy w większym lub mniejszym stopniu odnoszą się do sfery bezpieczeństwa, pozostają z nią w ścisłej korelacji.

Ważny jest też aspekt odnoszący się do bezpieczeństwa narodowego. O tym, czy dana dziedzina lub obszar powinny podlegać sferze tego bezpieczeństwa, nie mogą decydować w sposób kategoryczny wyłącznie argumenty prawne. W rzeczywistości bowiem konieczne jest uwzględnienie bieżącej sytuacji geopolitycznej oraz pozostałych, aktualnych czynników istotnych dla przedmiotowego zagadnienia.

Na potrzeby niniejszej pracy zdefiniowano bezpieczeństwo narodowe, jako brak zagrożeń dla norm i wartości narodu oraz brak obawy, że takowe zostaną naruszone. Zwiększoną uwagę poświęcono zwłaszcza bezpieczeństwu ekonomicznemu, szczególnie

[126] C. Znamierowski „Szkoła prawa. Rozważania o państwie", Warszawa 1999, s. 75 – 77.

w aspekcie personalnym. Naruszenie bezpieczeństwa wielu jednostek ma bowiem bezpośrednie przełożenie na bezpieczeństwo ogółu społeczeństwa, zwłaszcza tak szczególnego jak naród.

2.2 Deepfake – przegląd koncepcji badawczych

Kłamstwa informacyjne nie są niczym nowym. Jednak zdolność do zniekształcania rzeczywistości posunęła się gwałtownie naprzód dzięki technologii „głębokiego fałszu". Ta technologia umożliwia tworzenie audio i wideo prawdziwych ludzi, którzy mówią i robią rzeczy, których nigdy nie powiedzieli lub zrobili. Techniki uczenia maszynowego zwiększają stopień zaawansowania technologii, czyniąc głębokie podróbki coraz bardziej realistycznymi i odpornymi na wykrycie. Technologia *deepfake* ma cechy, które umożliwiają szybkie i szerokie rozpowszechnienie nagrań. Co więcej, oddając ją w ręce zarówno wyrafinowanych, jak i niewyszukanych aktorów, powoduje się możliwość niekontrolowanego i masowego wytwarzania takich materiałów. Wraz z dalszym rozwojem kart graficznych i poszerzeniem ich dostępności cenowej, wzrośnie liczba osób mogących w łatwy sposób wygenerować zmanipulowane treści.

Angielskie słowo *deepfake*, jest splotem dwóch innych angielskich wyrazów. *Deep* nawiązuje do *deeplearning*-u, czyli technologii głębokiego uczenia się (będącej formą sztucznej inteligencji) oraz *fake* – oszustwa, fałszerstwa. Słowo deepfake łączy więc w sobie terminy „głębokie uczenie się" i "fałszywe".

Powstanie pierwszych nagrań deepfake datuje się na koniec 2017 roku. Od tego czasu zarówno możliwości sprzętowe, jak i udostępnione oprogramowanie zostało wielokrotnie ulepszane. W 2018 roku możliwe było obrabianie twarzy w maksymalnej rozdzielczości 128 x 128 px. Rok 2019 i ulepszenie dostępnej technologii pozwoliły na edytowanie twarzy w rozdzielczości natywnej 224 x 224 px. Rok 2020 to pierwsze próby obróbki twarzy w rozdzielczości 384 x 384 px. oraz pod koniec roku, w rozdzielczości 448 x 448 px. Przełomowe było wówczas wprowadzenie przez Nvidie stosunkowo tanich i niezwykle popularnych kart z serii 30[127]. GeForce RTX 3090,

[127] Początkowo karty 3090 kosztowały po premierze ok. 7000 zł, co w porównaniu do dostępnych wcześniej, specjalistycznych kart Titan RTX, oferujących identyczną ilość pamięci GPU było bardzo dobrą okazją. W zależności od wersji, karty Titan RTX potrafiły wówczas kosztować od 12000 do 20000 zł. Co więcej, dzięki zastosowaniu nowej architektury w kartach z serii 30 – Ampere – zauważalnie zmalało zużycie energii względem dostępnej mocy. Ampere pozwoliło między innymi na asynchroniczne przyspieszone sprzętowo kopiowanie oraz zarządzanie

po dostosowaniu do niego bibliotek, pozwolił na edytowanie obrazów w natywnej rozdzielczości 512 x 512, co pozwoliło autorom nagrań na tworzenie pierwszych filmów w jakości full hd.

Przeprowadzona została analiza literatury w zakresie badań nad deepfake. Na przestrzeni ostatnich czterech lat przeprowadzono niewielką ilość badań obejmujących aspekty psychologiczne podatności na manipulację obrazem. Niewiele również dokonano prac bezpośrednio badających wpływ takowych nagrań na działania jednostki lub jej przekonania. Poniżej przedstawiono kilka najciekawszych z punktu niniejszej pracy wyników badań.

W 2020 roku Cristian Vaccari oraz Andrew Chadwick przeprowadzili badanie dotyczące podrabiania i dezinformacji: badanie wpływu syntetycznego filmu politycznego na oszustwo, niepewność i zaufanie do wiadomości[128]. Badanie przeprowadzone zostało na dużej, reprezentatywnej próbie populacji (N = 2005) Wielkiej Brytanii i pozwoliło badaczom porównać oceny ludzi dotyczące deepfake'ów. Odkryli oni, że ludzie częściej czują się niepewni niż wprowadzani w błąd przez deepfake, ale wynikająca z tego niepewność z kolei zmniejsza zaufanie do wiadomości w mediach społecznościowych. Wnioski, jakie wyciągnięte zostały z badania, sugerują, iż deepfake może przyczynić się do ogólnej nieokreśloności i cynizmu społecznego oraz spadku zaufania do materiałów audiowizualnych.

Wraz z pojawieniem się środków do tworzenia i rozpowszechniania fałszywych obrazów tworzonych z wykorzystaniem technologii deepfake, narodził się pomysł tworzenia zmanipulowanych nagrań seksualnych. Obrazy innych osób umieszczone zostały na nagraniach filmów pornograficznych w celu osiągnięcia korzyści finansowych, nękania lub zaspokojenia seksualnego swojej osoby. Trójka brytyjskich naukowców – Dean Fido, Jaya Rao, Craig A. Harper zauważyli potrzebę oceny i zrozumienia świadomości opinii publicznej i jej sądów na temat wspomnianego

pamięcią L2. Zastosowanie nowej technologii niosło jednak za sobą również duże wyzwania. Głównym z nich była konieczność utworzenia nowej wersji sterownika CUDA, przystosowaną dla tej serii kart. Wersję CUDA 11.1 wypuszczono dopiero pod koniec 2020 roku, a dostosowanie do niej biblioteki TensorFlow zajęło kolejne miesiące. Wersja beta oprogramowania DeepFaceLab dla kart z serii 30x pojawiła się dopiero w pierwszej połowie 2021 roku. Jeszcze pod koniec 2021 roku zdarzały się błędy związane między innymi z prawidłowym mergingiem nagrań.

[128] C. Vaccari, A. Chadwick, „Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News", social media + Society, 2020.

zachowania[129]. Przy dwóch, niezależnych próbach – zastosowali moderację i liniowe efekty mieszane, analizując czy osądy dotyczące deepfake różnią się w zależności od status ofiary (celebryta lub osoba niebędąca gwiazdą), danych demograficznych ofiary i uczestnika (wspólne pochodzenie, obszar zamieszkania), wykorzystania wizerunku (dzielenie się wizerunkiem w mediach społecznościowych) oraz własnego zadowolenia z życia seksualnego.

Badacze stwierdzili, że większą wyrozumiałością cieszą się autorzy deepfake, tworzący nagrania z celebrytami. Swoje poparcie dla nieszkodliwości takiego działania wyraziła zwłaszcza grupa męska, gdzie tworzenie nagrań usprawiedliwiane było ich wykorzystywaniem na użytek własny, a nie dzieleniem się nimi. Incydenty deepfake z udziałem ofiar – celebrytów były postrzegane jako mniej znaczące i szkodliwe przestępstwa w porównaniu z nagraniami dotyczącymi osób niebędącymi celebrytami. Co więcej, skłonność do łagodnych osądów, cechowała osoby z wyższą skłonnością do psychopatycznych zachowań. Badacze ustalili, iż w obu grupach badanych jedynie kolejno 6,6% i 4,4% osób, potrafiło dokładnie nazwać (lub zbliżyć się do nazwania) tworzenie i rozpowszechnianie fałszywej pornografii, utworzonej za pomocą *deepfake*.

W innym badaniu Pavela Korshunov oraz Sebastiena Marcel, udowodnili, iż dobrze wykonane nagranie deepfake jest w stanie rozpoznać jedynie 24,5% respondentów[130]. Badane osoby wiedziały jednak, że wśród prezentowanych im nagrań znajdują się zarówno nagrania prawdziwe, jak i fałszywe.

Jak dowodzą badacze z międzynarodowego zespołu Stefana Sütterlin, wpływającym na rozpoznanie fałszu wpływ może mieć również wielkość ekranu, na którym osoba ogląda dane wideo[131]. Respondenci oglądający filmy na telefonach udzielali błędnych odpowiedzi wielokrotnie częściej niż osoby oglądające i oceniające te same nagrania na dużym ekranie – tableta lub laptopa.

[129] D. Fido, J. Rao, C. A. Harper, „Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography", Judgements of Deepfake Media Production, 2020.
[130] P. Korshunov, M. Sébastien. „Deepfake detection: humans vs. machines", 2020.
[131] S. Sütterlin, T. F. Ask, S. Mägerle, Sophia, S. Gloeckler, et al. „Individual Deep Fake Recognition Skills are Affected by Viewers' Political Orientation, Agreement with Content and Device Used", 2021.

2.2.1 Rozwój technologii deepfake, a prawo

Z prawnego punktu widzenia powszechnie przyjmuje się, że prawo na całym świecie nie nadąża za rozwojem technologii deepfake, w tym wieloma sposobami wykorzystywania seksualnego obrazu[132]. Pomimo tego, w Japonii 2 października 2020 roku udało się przeprowadzić pierwsze na świecie aresztowanie osób odpowiedzialnych za tworzenie materiałów zniesławiających (pornograficznych) z wizerunkami znanych aktorek[133].

Obecnie w Europie nie ma żadnych wyraźnych informacji w ustawodawstwie dotyczących filmów deepfake. W Wielkiej Brytanii, gdzie sądy skupiają się wyłącznie na przestępstwach popełnionych z zemsty, zmanipulowane obrazy są objęte kluczowymi warunkami ochrony wizerunku. W Stanach Zjednoczonych nie istnieje żadne prawo federalne, które stanowiłoby podstawę prawną przeciwko deepfake'om. W 2019 roku pojawiły się jednak dwie inicjatywy Kongresu dotyczące materiałów deepfake. Pierwsza to zmuszenie Departamentu Bezpieczeństwa Wewnętrznego (Department of Homeland Security) do raportowania Kongresowi raportów dotyczących fałszerstw treści cyfrowych, takich jak filmy udostępniane z zamiarem wprowadzenia widza w błąd (Deepfake Report Act[134]). Ponadto ustawa zmuszałaby Departament do nałożenia ograniczeń na podmioty zagraniczne w zakresie kolportacji takich treści oraz dokonywania oceny dostępnych metod wykrywania i ograniczania zagrożeń tego typu.

Druga z inicjatyw (The Deep Fakes Accountability Act[135]) ma na celu penalizację tworzenia i dystrybuowania fałszywych wideo oraz ochronę ofiar i ograniczenie rozprzestrzeniania się cyfrowej dezinformacji. Dotychczas bowiem ofiary zmuszane były do korzystania z paragrafów zabraniających na przykład umieszczania wizerunków podobnych cyfrowo lub publikowania filmu pornograficznego bez uprzedniej zgody osoby nagrywanej. Zdarza się również, iż ofiary starają się ścigać twórców deepfake

[132] R. A. Delfino, „Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act", 88 Fordham L. Rev. 887, 2019.
[133] Adult video creation with „Deep Fake" technology or first Arrest", Japan Gazette.
[134] S. 2065 (116th): „Deepfake Report Act of 2019" – „an act „To require the Secretary of Homeland Security to publish an annual report on the use of deepfake technology, and for other purposes."
[135] H. R. 3230 – „Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability" Act of 2019.

z paragrafu „nękania" lub nagrywania części intymnych, lub biernego zaangażowania w akt seksualny[136].

Rebecca Delfino opisuje trudność w stwierdzeniu wiktymizacji osoby, której ciało znajduje się w deepfake'ach, Obecne ustawodawstwo nie oferuje żadnej ochrony jej ciała[137]. Ochrona ofiar deepfake zdaje się być niezwykle ważna, z racji mogących nieść ze sobą konsekwencji społecznych i rodzinnych, takich jak zniesmaczenie, rozpad relacji i zniszczona reputacja. Ofiara może również spotkać się z konsekwencjami zawodowymi, takimi jak wypowiedzenie zatrudnienia z powodu potencjalnego uszczerbku dla reputacji organizacji[138] oraz konsekwencji zdrowotnych, takich jak generowanie depresji, lęku i stresu związanego z utratą zaufania i rozpowszechnianiem obrazu siebie[139].

2.2.2 Elementy wpływające na siłę dezinformacji obrazem

Siła komunikacji wizualnej była klasycznym przedmiotem badań w badaniach nad komunikacją polityczną. W przełomowym eksperymencie Doris Graber odkryła, że widzowie telewizyjni z większym prawdopodobieństwem przypominają sobie komunikaty wizualne niż werbalne, a także na dłużej zapadają im one w pamięć[140]. Maria Grabe i Erik Bucy wykazali, że „gify" (tj. bezdźwięczne klipy, w których osoby są pokazywane, ale nie są słyszane) są silniejsze w kształtowaniu opinii wyborców niż samo mówienie[141]. Markus Prior wykazał zaś, że respondenci ankiet wykazywali wyższy poziom wiedzy, gdy pytania dotyczące zapamiętywania faktów zawierały zarówno informacje wizualne, jak i werbalne, co wskazuje na naturalną łatwość zapamiętywania tych treści[142].

Materiały wizualne poprawiają przekazywanie informacji, pomagając obywatelom tworzyć i odzyskiwać wspomnienia, odtwarzać ślad pamięciowy. Gabriella

136 D. Harris, „Deepfakes: False pornography is here and the law cannot protect you", Duke Law & Technology Review,17, 2019, s. 99-128.
137 R. A. Delfino, „Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act", 88 Fordham L. Rev. 887, 2019.
138 D. K. Citron, M. A. Franks, „Criminalizing Revenge Porn", Wake Forest Law Review, Vol. 49, 2014, p. 345+, U of Maryland Legal Studies Research Paper No. 2014-1.
139 S. Bates, „Revenge porn and mental health: A qualitative analysis of the mental health effects of revenge porn of female survivors", Feminist Criminology, 12(1), 2017 s. 22–42.
140 D. A. Graber, „Seeing is remembering: How visuals contribute to learning from television news", Journal of Communication, 40(3), 1990, 134–156.
141 M. E. Grabe, E. P. Bucy, „Image bite politics: News and the visual framing of elections", Oxford University Press, 2009.
142 M. Prior, „Visual political knowledge: A different road to competence?", Journal of Politics, 76(1), 2013, 41–57.

Stenberg wykazała, że jednostki przetwarzają informacje wizualne bardziej bezpośrednio i przy mniejszym wysiłku niż informacje werbalne[143]. Ilana Witten i Eric Knudsen argumentują, że ze względu na postrzeganą „precyzję" informacje wizualne są integrowane skuteczniej niż inne rodzaje danych sensorycznych[144]. Mylące wizualizacje są bardziej skłonne do generowania fałszywych wyobrażeń niż wprowadzające w błąd treści słowne, ponieważ w oparciu o "heurystykę realizmu", osoby traktują dźwięk i obrazy jako bardziej prawdziwe niż tekst[145]. Większa jest nieufność do treści tekstowych niż źródeł audiowizualnych, które z góry zakłada się za prawdziwe.

Kiedy obrazy i treści audiowizualne są łatwiejsze do zrozumienia i przetworzenia niż tekst pisany, w grę wchodzi „doświadczenie metapoznawcze" – wywiedzione doświadczalnie uczucia dotyczące naszego myślenia, które kształtują nasze reakcje na zadania, takie jak przetwarzanie nowych informacji[146]. Jedno z takich doświadczeń, „płynność" jest szczególnie ważne dla zrozumienia, dlaczego ludzie wierzą w fałszywe informacje. Ludzie są bardziej skłonni zaakceptować wiadomości jako prawdziwe, jeśli postrzegają je jako znajome[147]. Znajomość wywołuje „efekt prawdziwości" – poczucie płynności, które sprawia, że materiał jest łatwiejszy do przyswojenia, a przez to bardziej wiarygodny[148]. Ze względu na ich realizm techniczny, a zwłaszcza jeśli przedstawiają już znane osoby publiczne, fałszywe filmy (deepfake) polityczne mogą potęgować i tak już poważny problem polegający na tym, że płynność może być generowana poprzez znajomość, niezależnie od prawdziwości treści wideo.

Ważne jest również zachowanie użytkowników mediów społecznościowych w zakresie udostępniania. Filmy i zdjęcia są bardziej rozpowszechniane na portalu X (dawniej Twitter) niż wiadomości i petycje online[149]. Podczas kampanii prezydenckiej

143 G. Stenberg, „Conceptual and perceptual factors in the picture superiority effect", European Journal of Cognitive Psychology, 18(6), 2006, 813–847.
144 I. B. Witten, E. I. Knudsen, „Why seeing is believing: Merging auditory and visual worlds", Neuron, 48(3), 2005, 489–496.
145 S. J. Frenda, E. D. Knowles, W. Saletan, E. F. Loftus, „False memories of fabricated political events. Journal of Experimental Social Psychology", 49(2), 2013, 280–286.
146 N. Schwarz, L. J. Sanna, I. Skurnik, C. Yoon, „Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns", Advances in Experimental Social Psychology, 39, 2007, 127–161.
147 A. J. Berinsky, „Rumors and health care reform: Experiments in political misinformation". British Journal of Political Science, 47(2), 2017, 241–262.
148 E. J. Newman, M. Garry, C. Unkelbach, D. M. Bernstein, D. Lindsay, R. A. Nash, „Truthiness and falsiness of trivia claims depend on judgmental contexts", Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5), 2015, s. 1337–1348.
149 S. Goel, A. Anderson, J. Hofman, D. J. Watts, „The structural virality of online diffusion", Management Science, 62(1), 2015, s. 180–196.

USA w 2016 Tweety Donalda Trumpa i Hillary Clinton, które zawierały obrazy lub filmy, otrzymały znacznie więcej polubień i retweetów, niż tweety składające się wyłącznie z samej treści[150].

W ostatnich latach fałszywe wiadomości stały się problemem zagrażającym dyskursowi publicznemu, społeczeństwu ludzkiemu i demokracji[151]. Tak zwane „*fake news*-y" odnoszą się do fikcyjnych treści w stylu wiadomości, które są sfabrykowane w celu oszukania opinii publicznej[152]. Fałszywe informacje szybko rozprzestrzeniają się w mediach społecznościowych, gdzie mogą mieć wpływ na miliony użytkowników[153]. Obecnie jeden na pięciu internautów otrzymuje wiadomości za pośrednictwem YouTube, co w pozyskiwaniu informacji ustępuje wyłącznie Facebookowi, który jest w tym liderem. Ten wzrost popularności wideo podkreśla potrzebę tworzenia narzędzi do potwierdzania autentyczności treści medialnych i informacyjnych, ponieważ nowatorskie technologie pozwalają na przekonujące manipulowanie materiałem wideo[154]. Biorąc pod uwagę łatwość pozyskiwania i rozpowszechniania dezinformacji za pośrednictwem platform mediów społecznościowych, coraz trudniej jest powiedzieć, czemu ufać, a co skutkuje m.in. szkodliwymi konsekwencjami dla świadomego podejmowania decyzji[155].XXI wiek to epoka przez wielu nazywana „postprawdą". Charakteryzuje się dezinformacją cyfrową i wojną informacyjną prowadzoną przez złowrogich aktorów prowadzących fałszywe kampanie informacyjne w celu manipulowania opinią publiczną[156].

Niedawne postępy technologiczne ułatwiły również tworzenie tak zwanych deepfake, hiperrealistycznych filmów przy użyciu zamiany twarzy, które pozostawiają

150 E. Pancer, M. Poole, „The popularity and virality of political social media: Hashtags, mentions, and links predict likes and retweets of 2016 US presidential nominees' tweets", Social Influence, 11(4), 2016, s. 259–270.
151 L. Borges, B. Martins, P. Calado, „Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News", Journal of Data and Information Quality, 11(3): Article No. 14, 2019.
152 M. Aldwairi, A. Alwahedi, „Detecting Fake News in Social Media Networks", Procedia Computer Science, 141, 2018, s. 215–222.
153 A. Figueira, L. Oliveira, „The current state of fake news: challenges and opportunities", Procedia Computer Science, 121, 2017, s. 817–825.
154 K. E. Anderson, „Getting acquainted with social networks and apps: combating fake news on social media", Library HiTech News, 35(3), 2018, s. 1–6.
153 M. A. Britt, J. F. Rouet, D. Blaum, K. Millis, „A Reasoned Approach to Dealing with Fake News. Policy Insights from the Behavioral and Brain Sciences", 6(1), 2019, s. 94–101.
156 A. Qayyum, J. Qadir, M. U. Janjua, F. Sher, „Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News", IT Professional, 21(4), 2019, s. 16–24.

niewiele śladów manipulacji[157]. Deepfake to produkt aplikacji sztucznej inteligencji (AI), które łączą, zastępują i nakładają obrazy i klipy wideo w celu tworzenia fałszywych filmów, które wyglądają na autentyczne[158]. Technologia deepfake może na przykład wygenerować humorystyczny, pornograficzny lub polityczny film przedstawiający osobę mówiącą cokolwiek, bez zgody osoby, której wizerunek i głos jest używany[159]. Czynnikiem wpływającym na grę w przypadku deepfake jest zakres, skala i wyrafinowanie zastosowanej technologii, ponieważ prawie każdy posiadacz komputera może tworzyć fałszywe filmy, które są praktycznie nie do odróżnienia od autentycznych filmów[160]. Pierwsze przykłady deepfake'ów skupiały się na przywódcach politycznych, aktorkach, komikach i artystach estradowych, których twarze były nałożone na aktorów grających w filmach porno[161]. Deepfake w przyszłości prawdopodobnie będzie coraz częściej używany do zemsty, zastraszania, tworzenia fałszywego wideo jako dowodu w sądach, sabotaż polityczny, propaganda terrorystyczna, szantaż, manipulacja na rynku giełdowym i do przekazywania fałszywych wiadomości.

Wpływ technologii deepfake cały czas rośnie. Nie przeszkadza temu spadek zaufania do fotografii, które w przeciągu ostatnich kilkudziesięciu lat stopniowo podupada[162]. Nadal przykłada się bardzo dużą wagę do wszelkich dowodów fotograficznych czy filmowych[163]. Jeszcze większe zaufanie pokłada się w nagraniach głosowych, zwłaszcza jeżeli dany dźwięk jest nam znany[164]. System wzrokowy mózgu, mimo że jest w dużej mierze odporny na warunki naturalne, może być łatwym celem błędnego postrzegania. Klasyczne przykłady obejmują wszelkie złudzenia optyczne i niejednoznaczne rysunki (np. dobrze znany wazon Edgara Rubina, które można oglądać

157 R. Chawla, „Deepfakes: How a pervert shook the world", International Journal of Advance Research and Development, 4(6), 2019, s. 4–8.
158 M. H. Maras, A. Alexandrou, „Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos", International Journal of Evidence & Proof, 23(3), 2019, s. 255–262.
159 C. Day, „The Future of Misinformation", Computing in Science & Engineering, 21(1), 2019, s. 108–108.
160 J. Fletcher, „Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance", Project MUSE, Theatre Journal, 70(4), 2018, s. 455–471.
159 H. R. Hasan, K. Salah, „Combating Deepfake Videos Using Blockchain and Smart Contracts", IEEE Access, 7, 2019, s. 41596–41606.
162 J. Westling, „Deep Fakes: Let's Not Go Off the Deep End", 2019.
163 Y. Granot, E. Balcetis., N. Feigenson, T. Tyler, „In the eyes of the law: Perception versus reality in appraisals of video evidence", Psychology, Public Policy, and Law, 24(1), 2018, s. 93-104.
164 B. Brucato, „Policing made visible: Mobile technologies and the importance of point of View", Surveillance & society, 13(3/4), 2015, s. 455-473.

na dwa różne sposoby)[165]. Zaskoczenie i niedowierzanie związane z ”ujawnieniem sztuczek” potwierdzają, jak bardzo człowiek ufa swoim oczom – nawet jeśli wie, że zaraz zostanie oszukany. Jeśli osoba widzi coś na własne oczy, wierzy, że to istnieje lub jest prawdą, nawet jeśli jest mało prawdopodobne.

Rozpowszechnianie fałszywych informacji jest relatywnie łatwe, natomiast ich korekta oraz przeciwdziałanie dezinformacji, w tym fake news-om i deepfake-om, stanowi znacznie większe wyzwanie[166]. Skuteczna walka z tymi zjawiskami wymaga dogłębnego zrozumienia ich natury, przyczyn ich powstawania oraz technologii, które je umożliwiają. Warto podkreślić, iż badania naukowe dopiero niedawno zaczęły zajmować się dezinformacją multimedialną w mediach społecznościowych. Pierwsze filmy deepfake zadebiutowały w Internecie w październiku 2017 roku, co sprawia, że literatura naukowa w tym zakresie jest wciąż ograniczona. Niniejszy przegląd literatury ma na celu szeroką analizę tych zjawisk, w tym definicji filmów deepfake, profili ich twórców, a także potencjalnych korzyści i zagrożeń wynikających z tej technologii. Dodatkowo przedstawione zostały wybrane przykłady współczesnych wytworów deepfake oraz metody przeciwdziałania ich szkodliwym skutkom. W tym celu przeanalizowano szeroki zbiór artykułów informacyjnych oraz naukowych, a także materiały z konferencji dotyczących fake newsów i deepfake'ów.

2.2.3 Kategoryzacja materiałów tworzonych przy pomocy deepfake

W niniejszym podrozdziale zaprezentowane zostaną modele wykorzystania technologii deepfake w kreowaniu wirtualnej rzeczywistości. W punktach poniżej wymieniono obecne formy wykorzystania deepfake w edycji zdjęć, filmów, dźwięku oraz filmów z dźwiękiem. Do każdej kategorii podano przykład wykorzystania go oraz pomysł na pozytywne wykorzystanie technologii w biznesie.

Pod pojęciem deepfake w szerokim tego słowa znaczeniu kryje się nie tylko nakładanie obrazu jednej twarzy na drugą, lecz także inne techniki manipulacji obrazem. W poniższym zestawieniu umieszczono typologię ze względu na model wykorzystania głębokiego uczenia maszynowego oraz jego użycia w deepfake. Zestawienie zawiera opis oraz przykład aktualnego wykorzystania oraz wizję dalszego rozwoju i pomysły,

[165] T. C. Kietzmann, S. Geuter, P. König, „Overt visual attention as a causal factor of perceptual awareness”. PloS one, 6(7), 2011.
166 J. De keersmaecker, A. Roets, „ ‘Fake news’: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions", Intelligence, 65, 2017, s. 107–110.

w kierunku których dąży. Jako ostatni punkt każdego przykładu, prezentowane jest zagrożenie wynikające z dalszego rozwoju technologii oraz pomysły przeciwdziałania.

Jednym z możliwości manipulowania obrazem jest edycja zdjęcia z wykorzystaniem innego zdjęcia. Odbywa się to na zasadzie wyuczonych modeli i możliwe jest do wykonania zarówno poprzez aplikacje web-owe jak i aplikacje dostępne na telefon. Jednym z nich jest aplikacja FaceApp, której filtr starzenia zmienia zdjęcie w taki sposób, aby odmłodzić lub postarzyć prezentowaną na fotografii postać. W aplikacji możliwa jest również zarówno zamiana twarzy, jak i ciała – dokonywanie zmian w wyglądzie twarzy, zastępowanie lub mieszanie twarzy (lub ciała) z twarzą (lub ciałem) kogoś innego. Przy pomocy tej metody możliwe jest preparowanie fałszywego przekazu lub uwiarygodnianie fałszywych treści, jak również ośmieszenie danej osoby i niszczenie jej wizerunku.

Metoda ta, może być wykorzystywana zarówno przez oszustów, poprzez preparowanie w łatwy sposób zdjęć, jak i osoby zajmujące się dezinformacją. Możliwe jest preparowanie przy jej pomocy fałszywego przekazu lub uwiarygodnianie fałszywych treści. Dzięki tej metodzie możliwe jest również ośmieszenie danej osoby i niszczenie jej wizerunku.

Do pozytywnych aspektów potencjalnego wykorzystania gotowych modeli, należeć może personalizacja gotowych produktów lub usług. Sprzedawca stosując tę technologię, w łatwy sposób może zaproponować swoim klientom rozwiązanie, w którym będą oni mogli przymierzyć okulary, fryzury lub ubrania całkowicie wirtualnie, bez potrzeby zakładania ich. W takim schemacie konsument udostępniałby sklepowi jedynie swoje zdjęcie, dzięki czemu oferowane produkty dostosowywałyby się do jego sylwetki / wyglądu, celem dopasowania poszczególnych elementów.

Kolejnym rodzajem materiałów, przy których powstawaniu wykorzystywane jest głębokie uczenie maszynowe, jest trening audio. Rozróżniane są obecnie dwie metody – zmiany głosu osoby z jednego nagrania na inny lub przetwarzanie tekstu na głos.

Przy pierwszej metodzie, zamiana głosu lub naśladowanie czyjegoś głosu odbywa się poprzez dostarczenie próbki nagrania drugiej osoby i automatyczny trening klonujący oryginalny dźwięk. Najprawdopodobniej to właśnie korzystając z tej metody oszuści

w styczniu 2020, doprowadzili do podrobienia głosu prezesa firmy i wyłudzenia z niej 243000 dolarów[167].

Należy również wyróżnić pozytywne aspekty rozwoju tej metody. Algorytmy uczenia maszynowego mogą pozwolić utworzyć audiobooki z narracja brzmiącą w zależności od potrzeb – młodziej lub starzej, bardziej męsko lub kobieco. Głos może uwzględniać wówczas różne dialekty lub akcenty i dostosowywać się do odmiennych postaci. W znaczący sposób może to uatrakcyjnić audiobooki.

Drugą metodą treningu głosu, jest zamiana tekstu na mowę. Wpisując dowolny tekst i udostępniając nagrany głos, otrzymać możemy tekst czytany tym głosem. Jednym z pierwszych projektów ogólnodostępnych w Internecie była aplikacja web-owa notjordanpeterson.com, gdzie po wpisaniu tekstu otrzymywano nagranie wypowiedzi zbliżone brzmieniem do głosu Jordana B. Petersona, słynnego profesora psychologii, popularyzatora nauki.

Obecnie tworzenie fałszywych nagrań głosowych oferowane jest przez wiele platform, między innymi resemble.ai czy elevenlabs.io. Oba portale w swojej ofercie od lutego 2023 mają również język polski. Przyczyniło się to do powstania wielu nagrań wideo, na których pomieniony był język wypowiedzi osób na nich występujących. Filmiki takie dotknęły zarówno celebrytów, polityków czy również osoby związane z kościołem. Jest to jeden z aspektów, do których wykorzystywać można deepfake głosu – satyra. Większość tych nagrań miała bowiem charakter prześmiewczy.

Prognozuje się dalsze doskonalenie tej technologii i wykorzystanie jej również do celów między innymi politycznych. Pierwsze w Polsce, polityczne wykorzystanie deepfake głosu, opublikowane zostało przez Platformę Obywatelską 24.08.2023 roku na portalu X. Na nagraniu odczytywana jest, głosem przypominającym głos Prezesa Rady Ministrów Mateusza Morawieckiego[168], treść maila, który opublikowany został ze skrzynki mailowej Michała Dworczyka. Nagranie z początku nie zostało oznaczone jako wytworzone przy pomocy deepfake, stąd oglądający je mógł odnieść wrażenie, iż słowa te faktycznie wypowiada sam premier. Dopiero po kilku godzinach, z oficjalnego konta

[167] Akta opisujące sprawę oszustwa, https://www.documentcloud.org/documents/21085009-hackers-use-deep-voice-tech-in-400k-theft [dostęp: 01.01.2023].
[168] Profil portalu X z dostępnym nagraniem, https://twitter.com/Platforma_org/status/1694582809875062977?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1694582809875062977%7Ctwgr%5E49036a9506cfd80cfb4027224165552d9b44b33d%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fwww.money.pl%2Fgospodarka%2Fpo-w-spocie-wygenerowala-glos-premiera-przez-ai-strzal-w-stope-6934140360145856a.html [dostęp: 25.08.2023].

Partii, dodany został komentarz informujący o tym, iż głos nie jest prawdziwy, a wytworzony przez „technologię ai".

Pozytywnym aspektem rozwoju deepfake głosu jest możliwość wykorzystania go między innymi w kinematografii, gdzie w przypadku stwierdzenia błędnie wypowiedzianej sentencji lub konieczności zmiany słowa, swobodnie naprawić można głos aktora. Podobnie dzieje się w przypadku chęci wygenerowania automatycznego lektora, tłumaczącego film dowolnym głosem.

Ponadto edycja głosu i zmiana go na dowolny język, pozwala utworzyć nagrania, na których dany aktor wypowiada się w różnych językach. Zastosować można to nie tylko w kinematografii, tak by znany aktor posługiwał się biegle innymi językami, lecz również w reklamie czy filmach instruktażowych. Dzięki deepfake mogą być one „przetłumaczone" na inne języki, używając tego samego głosu, jaki użyty jest w oryginalnym nagraniu.

Oryginalnym tworem technologii deepfake jest jednak zmiana twarzy, poprzez wytrenowanie uniwersalnego obrazu osoby i nałożenie go na twarz osoby z innego nagrania. Obecnie popularnie stosuje się to w filmach. Jednym z pierwszych nagrań deepfake (2019 rok) była przeróbka programu Seth Meyers-a, gdzie twarz Jima Carreya zastępuje twarz występującej w nim Alison Brie[169].

W możliwości nakładania twarzy jednej osoby na drugą upatruje się wiele zagrożeń szczegółowo opisanych w niniejszej pracy. Obserwuje się próby podszywania pod znane osoby celem uprawiania dezinformacji, wpływu społecznego lub oszustwa. Dalszy rozwój technologii, a co za tym idzie i jakości nagrań spowodować może, iż powszechnie stosowana biometria rozpoznawania twarzy stanie się łatwa do oszukania.

Technologia deepfake może zostać użyta jednak również w pozytywnym kontekście. W 2023 roku utworzony został przy jej pomocy ostatni (589) odcinek serialu „Świat według Kiepskich", gdzie zmarłych aktorów – Dariusza Gnatowskiego i Ryszarda Kotysa – zastąpili Artur Kujawa i Michał Pietrzak[170]. Stosowanie deepfake upatruje się również jako metodę podmienienia twarzy głównego aktora na nagraniu z kaskaderem dla bardziej realistycznego ujęcia akcji w filmach.

[169] Przerobione nagranie wywiadu, https://youtu.be/b5AWhh6MYCg?si=qvbQeUjYoHi6UgGj [dostęp: 01.01.2023].
[170] 589 odcinek sezonu 31 Świata według Kiepskich, https://polsatboxgo.pl/wideo/seriale/swiat-wedlug-kiepskich/5024045/sezon-31/5034855/swiat-wedlug-kiepskich-odcinek-589/86e3bccd1ebd013e195ffa42022a19cf [dostęp: 01.05.2023].

Od 2 lat nagrania deepfake szeroko stosowane są w marketingu. Na rynku obecny jest szereg firm przetwarzających wizerunki osób, które wyraziły na to zgodę i udostępniają przy ich pomocy usługę tworzenia dowolnych nagrań. Jedną z takich firm jest hourone.ai, gdzie klient może zamówić nagranie danej twarzy wypowiadającej wymyśloną przez niego treść[171]. Dzięki temu osoba zlecająca może uzyskać profesjonalne nagranie dużo mniejszym kosztem niż wynajęcie profesjonalnego lektora użyczającego swój wizerunek, gdyż honorarium dla ochotnika jest przekazywane bezpośrednio przez firmę

Inną metodą zmian twarzy jest ich morfing. Technologia ta polega na tworzeniu płynnego przejścia z jednej twarzy na inną lub ewoluowaniu oryginalnej twarzy. Jednym z pierwszych zastosowań było wykorzystanie morfingu w filmie „Terminator II", gdzie twarze głównych aktorów zmieniają się w trakcie filmu na inaczej wyglądające. Inny przykład to przeróbka odcinka „Saturday Night Live", gdzie Bill Hader niepostrzeżenie zmienia się w Arnolda Schwarzeneggera[172]. Jedną z powszechnie stosowanych aplikacji do tego zabiegu jest facemorph.me[173]. Kolejnym zastosowaniem morfingu deepfake, jest możliwość tworzenia gier wideo, gdzie gracze będą mogli implikować w nich swoje twarze na postacie w grze i ewoluować je wraz z rozwojem postaci.

Innym zastosowaniem morfingu deepfake, jest możliwość tworzenia gier wideo, gdzie gracze będą mogli implikować w nich swoje twarze na postacie w grze i ewoluować je wraz z rozwojem postaci.

Pojęcie *Puppetry deepfake* również nie doczekało się tłumaczenia na język polski. Rozumieć je można jako „lalkowanie" deepfake, czyli tworzenie realnie wyglądających animacji twarzy lub całego ciała, na podstawie jednego zdjęcia. Jest to swoiste poruszanie daną postacią, jej elementami, podobnie jak z lalką. Możliwa jest również transpozycja ruchu ciała z jednej osoby na inną, poprzez nałożenie na nią swoistej maski. Nagranie „Everybody Dance Now" prezentuje, jak każdy może wyglądać niczym profesjonalny tancerz[174]. *Puppetry* powszechnie wykorzystywane jest przez aplikacje rozrywkowe dostępne na telefon, takie jak wombo.ai[175] czy reface.ai[176], które pozwalają stworzyć

[171] Strona startup-u Hour One, www.hourone.ai [dostęp: 01.01.2023].
[172] Przerobione nagranie wywiadu, https://youtu.be/bPhUhypV27w?si=DVqXcFRM0omKZhb2 [dostęp: 01.01.2023].
[173] Strona aplikacji MorphMe, https://facemorph.me/ [dostęp: 01.04.2023].
[174] Nagranie tańca stworzone przy pomocy puppetry deepfake https://youtu.be/PCBTZh41Ris?si=zP9qMHgnfbbctN20 [dostęp: 01.05.2023].
[175] Strona aplikacji Wombo, www.wombo.ai [dostęp: 01.06.2022].
[176] Strona aplikacji Reface, www.reface.ai [dostęp: 01.06.2022].

kilkunastosekundowy film, na którym wybrane przez nas zdjęcie porusza się w rytm muzyki lub nasz wizerunek umieszczony zostaje jako główny w krótkim fragmencie nagrania. Oprócz wykorzystania *puppetry* w marketingu sprzedażowym, istnieje również możliwość użycia go do budowania marki osobistej. Znany polityk, biznesmen, aktor czy sportowiec mogą ukryć swoje dolegliwości fizyczneczy urazy, podczas prezentacji wideo ukazujących ich w pełni sił.

Zwieńczeniem wszystkich technologii kryjących się lub powiązanych z pojęciem deepfake, jest ich połączenie w spójną całość. Dopiero prawidłowe zgranie audio, jak i wideo daje oglądającemu pełnię przekonania, iż ogląda prawdziwe nagranie. Prawidłowa synchronizacja ruchu warg i mimiki twarzy ze słowami wypowiedzianymi w danym nagraniu, pozwala na utworzenie doskonałego obrazu oszustwa, umożliwiającego podszywanie się pod dowolną osobę.

W nagraniu „You Won't Believe What Obama Says In This Video!" znany komik – Jordan Peele – użycza swojej twarzy do stworzenia fałszywego nagrania wypowiedzi Prezydenta USA Baracka Obamy[177]. Dzięki *puppetry* deepfake oraz fałszywemu nagraniu głosu metodą deepfake, możliwe jest odniesienie wrażenia, iż nagranie jest faktycznie fragmentem przemówienia Prezydenta.

Możliwości wykorzystania profesjonalnie wykonanych łączeń deepfake głosu z obrazem zdaje się być całe mnóstwo. Podobnie jak każda technologia, również i ta stwarza nowe zagrożenia jak i szanse. Deepfake pozwala bowiem kraść i nielegalnie wykorzystywać czyjś wizerunek, stwarzając zagrożenie nie tylko dla tej osoby, lecz również dla innych. W przypadku unormowania danych czynności i na przykład świadomego użyczenia swojego wizerunku do danego nagrania, pozwala jednak tworzyć piękne dzieła. Interesujące zjawisko obserwuje się wśród firm zajmujących się genealogią. Jedną z nich jest firma zajmująca się genealogią – MyHeritage, która pozwala przy pomocy technologii *deepfake* animować zdjęcia naszych bliskich krewnych, w tym zmarłych[178], dzięki czemu każdy może ożywić swoje wspomnienia. W perspektywie dalszego rozwoju, mając fragment nagranego głosu osoby zmarłej, możliwe będzie również wytrenowanie go i podkładanie pod dowolne filmy czy książki, tak by towarzyszył on osamotnionej osobie przez resztę czasu. Kierunkiem dalszego rozwoju

[177] Fałszywe nagranie przemówienia Prezydenta USA Baracka Obamy, https://youtu.be/cQ54GDm1eL0?si=G9NJwBDASTdP04HA [dostęp: 01.01.2021].
[178] Strona aplikacji MyHeritage, na której dostepna jest opcja „ożywiania" wspomnień, https://www.myheritage.pl/deep-nostalgia [dostęp: 01.10.2023].

może być zbieranie treści wiadomości z mediów społecznościowych oraz jej aktywności w Internecie, a następnie po poddaniu treningowi, wytworzeniu modelu sztucznej inteligencji, mającej wizerunek i głos osoby zmarłej i zachowującej się w zbliżony do niej sposób.

2.3 Psychologiczne aspekty podatności na dezinformację

Najbardziej obszerny, zróżnicowany i systematyczny dział badań nad dezinformacją stanowią prace z zakresu psychologii oraz nauk społecznych, wdrażając do analiz środowiska internetowego bogaty empiryczny i teoretyczny dorobek tych dyscyplin. Z punktu widzenia niniejszej dysertacji jest to nurt najbardziej przydatny: badacze koncentrują się na jednostce, wysiłki poznawcze ukierunkowują na wskazanie czynników i wyjaśnienie procesów odróżniania prawdy od fałszu. W zależności od badanego przedmiotu wyróżnić można trzy następujące główne nurty badawcze, skupione na analizie: stylu w jakim fałszywe wiadomości powstają (*style-based fake news analysis),* ich propagacji i rozprzestrzeniania *(propagation-based fake news analysis)* oraz zaangażowania użytkowników w tworzenie – świadome lub nieświadome propagowanie nieprawdziwych wiadomości *(user-based fake news analysis)*[179].

W ramach pierwszego z nurtów (*style-based fake news analysis*) analizy zogniskowane są na zagadnieniach struktury, technik uzyskiwania wiarygodności. Stosuje się tu dobrze empirycznie udokumentowane hipotezy z zakresu szeroko pojmowanej psychologii. Jest to na przykład hipoteza Udo Undeutscha zakładająca, że komunikaty prawdziwe i fałszywe różnią się od siebie istotnie zarówno pod względem treści, jak i jakości[180]. Model Undeutscha poddaje ocenie pięć elementów:

1) stałość relacji (utrzymywanie linii narracyjnej),

2) cechy czynności relacjonowania (język, wewnętrzna spójność wypowiedzi),

3) struktura osobowości świadka i poziom jego rozwoju,

4) motywacja świadka do składania zeznań,

5) cechy zawartości zeznania (pierwotne i wtórne kryteria oceny)[181].

[179] G. B, K. Abhishek, N. S. Achyuth, S. Dhananjay, P. B. Mrudula, „Determination of fake news using blockchain and IBM Watson", 2020, s. 2.
[180] U. Undeutsch, „Beurteilung der glaubhaftigkeit von aussagen", Handbuch der psychologie 11, 1967, s. 26–181.
[181] A. Pieszko – Sroka, „Czy zeznania są wiarygodne? Poszukiwanie metody ich oceny i rola psychologa w tym procesie", Przegląd Bezpieczeństwa Wewnętrznego 5/2011 s. 45.

Z kolei operacyjny postulat monitorowania rzeczywistości (*reality monitoring*) głosi, iż rzeczywiste (prawdziwe) zdarzenia charakteryzują się wyższą gęstością i jakością informacji o charakterze zmysłowo-percepcyjnym[182]. Prowadzone są także próby analiz komunikatów fałszywych pod kątem hipotezy, że kłamstwa wyrażane są inaczej pod względem elicytacji (sugerowania) oraz kontroli własnej (kontrola przekazu)[183]. Hipoteza ta zakłada, że osoby, które są zmuszone do kłamstwa w odpowiedzi na pytania lub sugestie, mogą wyrażać je w inny sposób niż te, które samodzielnie kontrolują sposób przekazu kłamstw.

Analiza fałszywych wiadomości zogniskowana na sposobie ich propagacji (*propagation-based fake news analysis)* wykorzystuje głównie modele epidemiologiczne. Akty dezinformacji postrzegane są na zasadzie reprodukujących się patogenów w określonych populacjach. Zjawiska te wyjaśniane są w kontekście dobrze ugruntowanych spostrzeżeń jak hipoteza odwrotnego efektu (*backfire effect)*, zwana również efektem rykoszetu. Głosi ona, iż dostarczone dowody przeciwne przekonaniom jednostki w pewnych warunkach wzmacniają pierwotne przekonania[184]. Podobną do niej jest hipoteza efektu oporu (*backlash effect*). Jest to zjawisko, które polega na tym, że ludzie stają się bardziej przekonani o swoich własnych przekonaniach po otrzymaniu dowodów, które obalają owe przekonania. Jest to forma dysonansu poznawczego, czyli psychicznego niepokoju, który ludzie odczuwają, gdy ich przekonania i zachowania są ze sobą sprzeczne. Oba efekty mogą występować w wielu różnych kontekstach, ale najczęściej są obserwowane w odniesieniu do przekonań politycznych, religijnych i naukowych.

Przykładem funkcjonowania efektu zwrotnego jest reakcja ludzi na zmiany klimatu. Mimo ogromnych dowodów na to, że działalność ludzka przyczynia się do zmian klimatu, niektórzy ludzie nadal zaprzeczają istnieniu tego zjawiska. W momencie, gdy są oni skonfrontowani z dowodami, które obalają ich przekonania, mogą stać się bardziej przekonani, że zmiany klimatu nie są rzeczywiste, zamiast zmienić swoje przekonania zgodnie z dowodami. Może to być spowodowane różnymi czynnikami, w tym naciskiem społecznym, by dostosować się do określonych przekonań,

[182] M. K. Johnson, C. R. Raye, „Reality monitoring", Psychological review, 88, 1, 1981, s. 67.
[183] M. Zuckerman, B. M. DePaulo, R. Rosenthal, „Verbal and Nonverbal Communication of Deception", In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 14, pp. 1-59). New York: Academic Press, 1981.
[184] B. Nyhan, J. Reifler, „When Corrections Fail: The Persistence of Political Misperceptions", Political Behavior 32(2), 2010, s. 303-330.

chęcią ochrony własnego ego oraz brakiem zrozumienia dowodów samych w sobie. Odwrotny efekt można również zaobserwować w odniesieniu do sposobu, w jaki ludzie reagują na przekazy polityczne. Na przykład, jeśli ktoś jest konfrontowany dowodami, które obalają jego przekonania polityczne, może stać się bardziej przekonany, że jego przekonania są prawdziwe, zamiast zweryfikować swoje stanowisko.

Podobnym zjawiskiem wyjaśniającym drogi propagacji jest efekt konserwatyzmu (*conservatism bias*) – jest to tendencja do niedostatecznego (niechętnego) korygowania przekonań w przypadku przedstawiania nowych, wiarygodnych dowodów[185]. Efekt konserwatyzmu, inaczej nazywany również inklinacją konserwatywną, to tendencja do utrzymywania obecnych przekonań i uprzedzeń, nawet jeśli są one niezgodne z nowymi informacjami czy dowodami. Jest to forma błędu poznawczego, która może prowadzić do trudności w przyjmowaniu nowych poglądów czy zmiany dotychczasowych przekonań. Efekt konserwatyzmu może wystąpić w różnych dziedzinach, takich jak polityka, nauka lub biznes. Może prowadzić do trudności w przyjmowaniu nowych rozwiązań czy technologii, a także utrudniać adaptację do zmieniających się warunków czy potrzeb.

Analizując styl fałszywych wiadomości wyróżnić można również zjawisko nazywane efektem zaprzeczenia lub od nazwiska autora – efektem Semmelweisa (*Semmelweis effect*). Opisuje on tendencję odrzucania nowych dowodów, które są one sprzeczne z ustalonymi normami i przekonaniami. Efekt Semmelweisa, inaczej zwany efektem ignorancji specjalisty, to zjawisko, które polega na tym, że ludzie są niechętni do przyjmowania nowych pomysłów lub teorii, jeśli są one sprzeczne z ich dotychczasowymi przekonaniami czy wiedzą. Jest to forma błędu poznawczego, która może prowadzić do oporu przed zmianą dotychczasowych przekonań czy nawet do ich utrwalania.

Efekt Semmelweisa został nazwany na cześć Ignacego Semmelweisa, lekarza z XIX wieku, który stwierdził, że zakażenia szpitalne są spowodowane przez mikroorganizmy i można im zapobiegać poprzez dezynfekcję rąk. Jego teoria była sprzeczna z panującymi wówczas przekonaniami, według których zakażenia szpitalne były spowodowane „miazmatami" (gr. *miasma,* czyli splamienie, brud*)* lub innymi niezidentyfikowanymi czynnikami takimi jak niezdrowe powietrze czy zaduch. Pomimo

[185] S. Basu, „The conservatism principle and the asymmetric timeliness of earnings", Journal of Accounting and Economics, volume 24, 1997, s. 3 – 37.

tego, że Semmelweis dostarczył dowodów na poparcie swojej teorii, jego pomysły były ignorowane przez większość lekarzy, którzy byli przekonani o swojej wiedzy i nie byli skłonni zmienić swoich przekonań. Efekt Semmelweisa może być szczególnie widoczny w dziedzinie nauki, gdzie nowe teorie czy pomysły muszą przejść przez proces recenzji i akceptacji, aby być uznane za prawdziwe.

Analiza fałszywych wiadomości zogniskowana na ich użytkowniku i jego zachowaniu *(user-based fake news analysis*) skupia się na sposobie angażowania odbiorców w fałszywe wiadomości i ich roli w tworzeniu, rozpowszechnianiu lub ich odkrywaniu i zgłaszaniu. W analizach tych wyróżnia się grupę tzw. złośliwych użytkowników (*malicious users*), którzy celowo tworzą i/lub rozpowszechniają fałszywe wiadomości motywowani korzyściami natury ideologicznej i/lub ekonomicznej oraz zwykłych użytkowników (*normal users*), spośród których niektórzy rozpowszechniają fałszywe wiadomości wraz ze złośliwymi użytkownikami, powodowani niewiedzą, naiwnością, wpływem własnym lub efektem wpływu społecznego[186].

Analiza *fake news* oparta na użytkowniku to podejście, które polega na wykorzystaniu informacji o użytkownikach, którzy dzielą się nieprawdziwymi informacjami, do wykrywania i zwalczania tego typu treści. Jedną z jego metod jest analiza sieci społecznościowych, polegająca na zidentyfikowaniu użytkowników, którzy dzielą się nieprawdziwymi informacjami i śledzeniu, jak te informacje są rozpowszechniane. Możliwe jest przy tym wykorzystanie narzędzi do analizy języka naturalnego, takich jak uczenie maszynowe (*machine learning)* czy szerzej narzędzi AI, służących do automatycznego wykrywania nieprawdziwych informacji i odróżniania ich od prawdziwych. Pozwala to na identyfikowanie kont pod względem ich prawdziwości oraz budowania sieci wzajemnych powiązań. Analiza fałszywych wiadomości zogniskowana na użytkowniku ma kilka zalet. Po pierwsze, pozwala zidentyfikować i zablokować użytkowników, którzy dzielą się nieprawdziwymi informacjami, co może zapobiegać ich dalszemu rozprzestrzenianiu. Po drugie, pozwala zrozumieć, jak *fake newsy* są rozpowszechniane i jakie mechanizmy stoją za ich dystrybucją, co może pomóc w opracowaniu skutecznych strategii do ich zwalczania. Należy jednak zaznaczyć, iż analiza fake news użytkownika ma swoje ograniczenia. Może być trudna do zastosowania w przypadku treści, które są rozpowszechniane w sposób anonimowy lub

[186] Yuanbo Xu, Yongjian Yang, En Wang, Fuzhen Zhuang, Hui Xiong, „Detect Professional Malicious User with Metric Learning in Recommender Systems", Journal of Latex Class Files, Vol. 14, No. 8, 2020.

za pośrednictwem różnych kont czy stron. Ponadto, istnieje ryzyko nadużycia takich narzędzi przez rządy lub inne podmioty, które mogą wykorzystywać je do cenzurowania lub kontrolowania treści, które są uważane za niepożądane. Dlatego ważne jest, aby takie narzędzia były stosowane zgodnie z zasadami etyki i zapewniały transparentność oraz odpowiednie mechanizmy ochrony praw użytkowników.

Poruszając temat dezinformacji przez pryzmat zjawisk psychologicznych, należy poświęcić szczególną uwagę temu jak nasze myśli, uczucia i przekonania wpływają na nasze zachowanie. Efekt wpływu własnego *(self-influence effect)* obejmuje licznie empirycznie potwierdzone generalizacje empiryczne. Jest to między innymi efekt potwierdzenia (*confirmation bias*). Efekt potwierdzenia, to tendencja do szukania i przyjmowania informacji, które potwierdzają nasze dotychczasowe przekonania lub uprzedzenia, a ignorowania lub odrzucania tych, które są sprzeczne z nimi. Jest to forma błędu poznawczego, która może prowadzić do utrwalania błędnych przekonań czy uprzedzeń oraz trudności w przyjmowaniu nowych pomysłów lub teorii[187]. Jego konsekwencją mogą być nieobiektywna ocena informacji czy argumentów, błędne wnioski i decyzje.

Jednym z powodów, dla których ludzie mogą mieć skłonność do tendencji potwierdzenia, jest lęk przed zmianą lub osłabianie motywacji do zmiany. Zmiana dotychczasowych przekonań może być postrzegana jako trudna lub niebezpieczna, co prowadzi do oporu przed przyjmowaniem nowych informacji czy pomysłów. Innym powodem jest brak czasu lub zasobów na poświęcenie uwagi nowym informacjom czy nauczenie się nowych umiejętności[188]. Ważnym aspektem jest również fakt, iż lubimy czuć się dobrze i mieć rację. Efekt potwierdzenia poprzez pozytywne wzmacnianie sprawia, że czujemy się dobrze, ponieważ potwierdzane są nasze przekonania, pomaga również zredukować efekt dysonansu poznawczego[189].

Innym zjawiskiem psychologicznym jest iluzja asymetrycznego wglądu (*illusion of asymmetric insight*), zwana również iluzją asymetrycznej wiedzy. Jest to skłonność do przeceniania swojej wiedzy i umiejętności rozumienia, na podstawie uznawania

[187] R. S. Nickerson, „Confirmation Bias: A Ubiquitous Phenomenon in Many Guises", Review of General Psychology 1998, Vol. 2, No. 2, s. 175-220.
[188] B. Dardenne, J. P. Leyens. „Confirmation Bias as a Social Skill", „Personality and Social Psychology Bulletin". 21 (11), 1995, s. 1229–1239.
[189] R. S. Nickerson, „Confirmation bias: A ubiquitous phenomenon in many guises", Review of General Psychology, 2(2), 1998, s. 175–220.

swojej wiedzy i umiejętności za przewyższające wiedzę oraz umiejętności innych[190]. Iluzja asymetrycznej wiedzy to zjawisko, które polega na przekonaniu, że mamy więcej wiedzy lub lepsze zrozumienie pewnej sytuacji niż inni, co może prowadzić do zniekształceń poznawczych.

Iluzja asymetrycznej wiedzy może mieć szczególne znaczenie w sytuacjach, gdy podejmujemy decyzje zespołowe lub negocjujemy. Może prowadzić do braku szacunku dla innych osób i ich poglądów oraz do niedoceniania ich wiedzy i doświadczenia, współistniejąc z brakiem zrozumienia, dlaczego inni ludzie mogą mieć inne podejście do danej sytuacji czy odmienne priorytety[191].

Do metod wykorzystywanych w badaniach efektów należy także naiwny realizm (*naive realism*). Jest to przekonanie, iż zmysły zapewniają nam bezpośrednią świadomość przedmiotów takimi, jakimi naprawdę one są[192]. Naiwny realizm to filozoficzny pogląd, według którego nasze doświadczenia bezpośrednio odzwierciedlają rzeczywistość. Oznacza to, że ludzie postrzegają świat taki, jaki jest rzeczywiście, bez dodatkowej interpretacji czy filtracji przez myśli lub emocje. Naiwny realizm jest zazwyczaj kojarzony z prostym, nieanalizującym sposobem myślenia, który nie uwzględnia kontekstu ani interpretacji okoliczności.

Ten pogląd jest często krytykowany przez inne teorie filozoficzne, takie jak subiektywizm czy idealizm, które twierdzą, że nasze doświadczenia są w jakiś sposób zniekształcone przez nasze indywidualne przekonania i emocje czy kontekst kulturowy. W odróżnieniu od tych poglądów, naiwny realizm twierdzi, że nasze doświadczenia w codzienności są w pełni obiektywne i odzwierciedlają rzeczywistość taką, jaka jest w ten sposób teoria naiwnego realizmu nie uwzględnia tego, iż doświadczenia mogą być ograniczone przez zmysły[193].

Efekt nadmiernej pewności siebie (*overconfidence effect*) to subiektywne zaufanie danej osoby do jej osądów, które jest zdecydowanie większe niż obiektywne[194]. Efekt

---

[190] E. Pronin, J. Kruger, K. Savitsky, R. Kenneth Lee. „You Don't Know Me, But I Know You: The Illusion of Asymmetric Insight", Journal of Personality and Social Psychology – PSP. 81. 2001, s. 639-656.
[191] A. Steglich-Petersen, M. Skipper, „Explaining the Illusion of Asymmetric Insight". Review of Philosophy and Psychology 10 (4), 2019, s. 769-786.
[192] A. Ward, L. Ross, E. Reed, E. Turiel, T. Brown, „Naive realism in everyday life: Implications for social conflict and misunderstanding". Values and Knowledge, 1997, s. 103-135.
[193] K. Adamska, „Iluzja transparentności – przyczyny i skutki", Studia Psychologiczne, t. 50 z. 4, 2012, s. 13 – 25.
[194] D. Dunning, D. W. Griffin, J. D. Milojkovic, L. Ross, „The overconfidence effect in social prediction", Journal of Personality and Social Psychology, 58(4), 1990, s. 568–581.

nadmiernej pewności siebie jest zjawiskiem, w którym ludzie są skłonni do przeceniania swojej wiedzy, umiejętności lub szans na sukces. Może to prowadzić do nieodpowiednio wysokiej samooceny lub nadmiernego ryzyka w działaniu. Zjawisko to jest szeroko badane w psychologii i ekonomii, ponieważ ma istotne konsekwencje dla decyzji podejmowanych przez ludzi.

Z efektem nadmiernej pewności siebie blisko związany jest efekt Dunninga-Krugera (*Dunning-Kruger effect*). Jest to błąd poznawczy, mówiący o tym, iż osoby o niskim poziomie wiedzy lub umiejętności w jakiejś dziedzinie mają tendencję do przeceniania swoich możliwości, podczas gdy osoby o wysokim poziomie wiedzy lub umiejętności mają tendencję do niedoceniania swoich możliwości[195]. Efekt ten najczęściej tłumaczy się brakiem metakognicji u osób o niskim poziomie wiedzy lub umiejętności. Nie mają one wystarczająco dużo kompetencji, aby móc ocenić swoją wiedzę lub umiejętność w sposób obiektywny. Odmienna sytuacja dotyczy osób z wysokim poziomem umiejętności lub wiedzy. Są one świadome swoich możliwości, natomiast porównując się do innych, mogą nie doceniać swoich osiągnięć. Istnienie efektu Dunninga-Krugera jest jednak negowane przez część środowiska psychologicznego[196]. W literaturze akademickiej toczy się na ten temat dyskusja[197],[198].

Z kolei wpływ społeczny wyraża się w następujących hipotetycznie zachodzących powszechnie efektach: inklinacji do ogniskowania uwagi konsumenta informacji na elementach, które w sposób ciągły i systematyczny docierają do jego świadomości (tzw. *attentional bias*)[199]. Attentional bias to zjawisko, w którym nasze zachowanie i postrzeganie są ukierunkowane przez to, na co skupiamy naszą uwagę. Może to prowadzić do preferencji lub uprzedzeń, ponieważ nasz mózg jest bardziej skłonny do przetwarzania informacji, na których częściej skupiamy uwagę.

---

[195] J. Kruger, D. Dunning, „Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments", Journal of Personality and Social Psychology, 77 (6), 1999, s. 1121–1134

[196] P. Juslin, A. Winman, H. Olsson, „Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect", Psychological Review, 107 (2), 2000, s. 384–396.

[197] J. Krueger, Ross A. Mueller, „Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance", Journal of Personality and Social Psychology, 82 (2), 2002, s. 180–188,

[198] G. E. Gignac, M. Zajenkowski, „The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data", Intelligence, 80, 2020, a. 101449.

[199] C. MacLeod, A. Mathews, P. Tata, „Attentional bias in emotional disorders", Journal of abnormal psychology 95, 1, 1986, s. 15.

Istnieją różne rodzaje uwagi, takie jak uwaga automatyczna, uwaga celowa i uwaga odwracalna. Uwaga automatyczna to proces, w którym nasza uwaga jest skierowana na bodźce, które są ważne dla nas lub są zgodne z naszymi przekonaniami lub oczekiwaniami, bez naszej świadomości. Uwaga celowa to proces, w którym nasza uwaga jest świadomie skierowana na określone bodźce lub zadania. Uwaga odwracalna to proces, w którym nasza uwaga jest skierowana na bodźce lub zadania, ale możemy ją łatwo przenieść na inne rzeczy. Z innej perspektywy możemy mówić o tendencji uwagi, czyli skłonności do skupiania się na pewnych typach informacji, uprzedzenia uwagi, co może prowadzić do tego, że będziemy faworyzować pewne informacje i ignorować inne. Ponadto skupienie uwagi pozwala koncentrować się świadomie na danym elemencie.

Istnieje wiele czynników, które mogą wpływać na naszą uwagę, takich jak emocje, przekonania i oczekiwania. Jeżeli jesteśmy zaniepokojeni lub zastraszeni, możemy być bardziej skłonni do skupienia się na bodźcach lub sytuacjach, które są związane z naszym zaniepokojeniem lub strachem. Podobnie, jeśli mamy pewne przekonania lub oczekiwania dotyczące danej sytuacji, możemy być bardziej skłonni do skupienia się na bodźcach, które są zgodne z naszymi przekonaniami lub oczekiwaniami.

Efekt wiarygodności (*validity effect*) to skłonność jednostek do dawania wiary wielokrotnie powtarzanym informacjom[200]. Jest to zjawisko, w którym wielokrotna ekspozycja na informację prowadzi do wzrostu postrzeganej wiarygodności tej informacji. Innymi słowy, im częściej coś widzimy lub słyszymy, tym bardziej prawdopodobne jest, że uwierzymy, że to prawda, nawet jeśli nie jest to do końca dokładne. Efekt ten jest blisko związany z pojęciem znajomości. Kiedy wielokrotnie słyszymy daną informację, staje się ona znajoma. Znajomość z kolei może prowadzić do uczucia prawdy lub komfortu, szczególnie jeśli informacja jest przedstawiana w sposób przekonujący.

Zbliżonym efektem jest zjawisko „zasada podczepienia" (*bandwagon effect*), czyli zjawisko niższego oporu dla powtarzania czynności i poglądów demonstrowanych przez innych masowo i publicznie[201]. Efekt ten powoduje, iż bardziej prawdopodobne jest, że ludzie podejmą określoną decyzję lub wykonają określoną czynność, jeśli widzą,

[200] L. E. Boehm, „The validity effect: A search for mediating variables", Personality and Social Psychology Bulletin 20, 3, 1994, s. 285–293.
[201] H. Leibenstein, „Bandwagon, snob, and Veblen effects in the theory of consumers' demand", The quarterly journal of economics 64, 2, 1950, s. 183–207.

że robią to inni. Innymi słowy, ludzie mają tendencję do ”podłączania się” do tłumu, ponieważ chcą być częścią czegoś popularnego lub chcą robić to, co robi większość.

Zasada podczepiania nazywana jest również niekiedy efektem mody lub efektem płynięcia z prądem, efektem owczego pędu. Angielskie „*bandwagon*” to nic innego jak wóz z muzyką, za którym na paradach, protestach, wiecach, podążają tłumy. Ludzie są skłonni do przyjmowania pewnych opinii lub podejść, ponieważ są one uważane za powszechne lub popularne. Może to prowadzić do zmiany naszych przekonań lub zachowań, aby dostosować się do większości lub do oczekiwań grupy. Efekt podczepiania może prowadzić do utrwalania błędnych lub niepełnych przekonań, a także do trudności w podejmowaniu obiektywnych i racjonalnych decyzji, braku krytycyzmu wobec podejść lub opinii, które są powszechne lub popularne.

W obszarze dezinformacji powszechnie uznaje się efekt iluzorycznej prawdy (*illusory truth effect*)[202]. Jest to zjawisko tożsame z dwoma wcześniej omówionymi i zakłada, iż fałszywa informacja wielokrotnie powtórzona staje się prawdziwa. Oceniając prawdę, ludzie często polegają na tym, czy informacja wydaje im się znajoma. Badacze wskazują, iż powtarzanie zwiększa wiarę w dezinformację, taką jak fałszywe nagłówki wiadomości i teorie spiskowe. Co istotne, wiarygodność rośnie nawet w przypadku twierdzeń nieprawdopodobnych lub sprzecznych z wcześniejszą wiedzą osób badanych. Ponadto, oprócz wpływu na wiarygodność, powtarzanie ma również szersze konsekwencje, takie jak zwiększanie chęci udostępniania nagłówków informacyjnych i zmniejszanie postrzegania nieetyczności danego czynu[203].

Zjawiskiem powszechnie występującym w mediach społecznościowych jest efekt komory echa (*echo chamber effect*). Efekt ten zakłada, iż obecnie ludzie otoczeni są informacjami, które potwierdzają ich istniejące przekonania i poglądy, jednocześnie ograniczając ekspozycję na inne punkty widzenia. Powoduje to ciągłe wzmacnianie wiary w informację w zamkniętych systemach informacyjnych, to jest względnie lub bezwzględnie izolowanych informacyjnie grupach[204]. Na efekt komory echa składają się głównie trzy elementy: algorytmy, sieci społeczne oraz potwierdzenie oczekiwań. Platformy mediów społecznościowych wykorzystują algorytmy, które personalizują

[202] L. Hasher, D. Goldstein, T. Toppino, „Frequency and the conference of referential validity”, Journal of Verbal Learning and Verbal Behavior, Volume 16, Issue 1, 1977, s. 107-112.
[203] J. Udry, S. J. Barber, „The illusory truth effect: A review of how repetition increases belief in misinformation”, Current Opinion in Psychology, Volume 56, 2024.
[204] K. H. Jamieson, J. N. Cappella, „Echo chamber: Rush Limbaugh and the conservative media establishment”, Oxford University Press; The Spreading of Misinformation Online, 2008.

treść, którą widzą użytkownicy. Oznacza to, że użytkownicy są bardziej narażeni na treści, które lubią, angażują się z nimi lub które algorytm uważa za interesujące na podstawie ich wcześniejszej aktywności. Ponadto ludzie zazwyczaj łączą się z osobami o podobnych poglądach, co również dzieje się w wirtualnym świecie. Prowadzi to do sytuacji, w której ludzie są otoczeni głównie przez osoby, które lubią i które potwierdzają ich przekonania.

W kształtowaniu społecznej świadomości niezwykle istotny jest efekt kaskady dostępności (*availability cascade*). Jest to samowzmacniający się proces kształtowania zbiorowych przekonań. Działa on poprzez wzajemne oddziaływanie tego, jak łatwo dostępne są informacje w przestrzeni publicznej oraz tego, jak te informacje wpływają na nasze postrzeganie prawdopodobieństwa i powszechności danego zjawiska[205]. Ludzie są bardziej skłonni do uwzględniania informacji, które są łatwo dostępne lub widoczne, ignorując informacje, które są trudniej dostępne lub mniej widoczne. Przekłada się to często do powstania błędnych lub niepełnych przekonań, a także do trudności w podejmowaniu obiektywnych i racjonalnych decyzji. Kaskada dostępności może być szczególnie widoczna w przypadku mediów, ponieważ informacje, które są często powtarzane lub zamieszczane w widocznych miejscach, są bardziej dostępne dla ludzi i mogą mieć większy wpływ na ich przekonania i zachowanie. Może to prowadzić do tworzenia się nieprawdziwych lub niepełnych narracji, które są potem przekazywane dalej.

Analogiczne wnioski sugerują Bayesowskie teorie przetwarzania informacji. W ramach tego nurtu formułuje się twierdzenia, jakoby jednostki aktualizowały swoje stanowiska polityczne w odpowiedzi na nowe informacje, w kierunku zgodnym z ich dotychczasowymi poglądami, postawami i wiedzą[206]. W literaturze przedmiotu wykazano wspomniany na wstępnie efekt rykoszetu (*backfire effect*) nazywany niekiedy efektem bumerangu (*bumerang effect*). Zgodnie z tym założeniem próba skorygowania błędnych przekonań u kogoś może wzmocnić te przekonania. Próba przekonania kogoś, że się myli, może sprawić, że będzie trzymał się swoich błędnych przekonań jeszcze bardziej kurczowo. Jest to zjawisko, zgodnie z którym ludzie są bardziej skłonni do utrwalania swoich przekonań lub opinii, gdy są one zagrożone lub zakwestionowane.

[205] T. Kuran, C. R. Sunstein, „Availability cascades and risk regulation", Stanford Law Review, 1999, s. 683–768.
[206] J. G. Bullock, „Partisan Bias and the Bayesian Ideal in the Study of Public Opinion", The Journal of Politics 71 (3), 2009, s. 1109-1124.

Powodem takiego stanu rzeczy jest kilka współzależnych od siebie reguł. Reaktancja zakłada ciągłe dążenie do przywrócenia wolności wyboru danej osoby zagrożonej przez inną osobę próbującą jej coś narzucić lub czegoś zakazać. Ponadto heurystyka potwierdzenia sugeruje, iż ludzie na podstawie istniejącej wiedzy stawiają hipotezy, których nie widzą sensu udowadniać, natomiast heurystyka dostępności powoduje, iż ludzie są bardziej skłonni wierzyć w informacje, które są im znajome. Oprócz tego polaryzacja grupowa społeczeństwa doprowadza do tego, iż próba skorygowania błędu u kogoś może być postrzegana jako atak na jego grupę. Zjawisko to nasila się wraz z wiekiem[207].

W kontekście zwalczania dezinformacji zjawisko to zostało przebadane między innymi przez Ullricha Eckera[208], który eksperymentami weryfikował czy pojawienie się sprostowań fałszywych informacji, wraz z powtórzeniem dezinformacji wpływa w pewnych warunkach na ich utrwalanie. Korekta, która przedstawiała uczestnikom nowe fałszywe informacje, nie prowadziła do silniejszych błędnych przekonań w porównaniu z grupą kontrolną, która nie była narażona na fałszywe informacje ani na korektę. Wyniki sugerują, że powtarzanie dezinformacji podczas jej korygowania jest bezpieczne, nawet gdy odbiorcy nie są wcześniej zaznajomieni z tą dezinformacją.

Efekt motywowanego rozumowania (*motivated reasoning bias*) powiązany jest ściśle z motywowanym poznaniem (*motivated cognition*)[209]. Zjawisko to zakłada, iż procesy myślowe i wnioskowanie ludzi są w dużym stopniu kontrolowane przez nasze pragnienia, cele, emocje. Nie zawsze myślimy w sposób obiektywny i logiczny, a rozumowanie jest mocno zabarwione motywacjami i przekonaniami. Efekt ten motywuje ludzi do tego, żeby analizować bardziej idee, które im się podobają, niż te, z którymi się nie zgadzają. Ludzie przetwarzają informacje w sposób, który potwierdza ich istniejące przekonania lub preferencje.

Motywowane rozumowanie może być szczególnie widoczne w przypadku kontrowersyjnych lub emocjonalnie ważnych kwestii, gdy ludzie są bardziej skłonni do obrony swoich przekonań. Może również być związane z poczuciem zagrożenia lub

[207] G. Pennycook, D. G. Rand, „The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings", Management Science, Vol 66(11), 2020, s. 4944-4957.
[208] Ecker, U.K.H., Lewandowsky, S. & Chadwick, M. Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. Cogn. Research 5, 41, 2020.
[209] C. S. Taber, M. Lodge, „Motivated skepticism in the evaluation of political beliefs". American Journal of Political Science 50 (3), 2006, s. 755-769.

ograniczenia swobody wyboru. Jest to również poniekąd związane z redukcją dysonansu poznawczego. Gdy ktoś trzyma się jednocześnie dwóch sprzecznych przekonań lub gdy jego przekonania są niezgodne z jego działaniami, motywowane rozumowanie pomaga zmniejszyć ten dyskomfort, znajdując sposoby na usprawiedliwienie tych działań lub porzucając niewygodne informacje.

Kolejnym zagadnieniem wpływającym na sądy poznawcze jest efekt walencji (*valence effect).* Jest to zjawisko, zgodnie z którym ludzie mają tendencję do oceniania informacji bardziej pozytywnie, jeśli jest ona powiązana z pozytywnymi bodźcami i bardziej negatywnie, jeśli jest powiązana z negatywnymi bodźcami. Ludzie są bardziej skłonni do przyjmowania pozytywnych informacji niż negatywnych, lub odwrotnie[210].

Związane jest to blisko z wgranymi asocjacjami, gdyż ludzie automatyczne tworzą skojarzenia między bodźcami. Kiedy widzimy coś pozytywnego, na przykład szczęśliwą twarz, wywołuje to pozytywne uczucia (przypisujemy pozytywną walencję). Z czasem możemy zacząć kojarzyć pozytywne emocje z samą twarzą, niezależnie od jej wyrazu. Ponadto heurystyka afektywna wskazuje, iż wykorzystujemy nasze emocje jako wskazówkę do wydawania szybkich osądów. Jeśli coś wywołuje w nas pozytywne uczucia, jesteśmy bardziej skłonni je zaakceptować, nawet jeśli nie mamy na to racjonalnych dowodów.

Rozwinięcie tej teorii mówi, iż ludzie przeceniają prawdopodobieństwo wystąpienia dobrych rzeczy, a nie doceniają prawdopodobieństwa zdarzenia się rzeczy złych[211]. Jest to tak zwane myślenie życzeniowe (*wishful thinking*). Wpływa ono na kształtowanie przekonań i podejmowanie decyzji zgodnie z tym, co może być przyjemne do wyobrażenia, zamiast odwoływać się do dowodów lub racjonalności[212]. Gdy mamy silne nadzieje na coś, zaczynamy wierzyć, że to się naprawdę wydarzy, nawet jeśli szanse na to są niewielkie lub żadne. Oprócz tego, że myślenie życzeniowe jest błędem poznawczym i nieefektywnym sposobem podejmowania decyzji, to może być również specyficznym błędem logicznym w argumentacji.

Ostatnią, bardzo dobrze opisaną w literaturze teorią jest teoria perspektywy (*prospect theory*). Powstała na styku nauk psychologicznych i ekonomicznych i zakłada,

[210] D. L. Rosehan, S. Messick, „Affect and expectation", Journal of Personality and Social Psychology 3, 1966, s. 38-44.
[211] N. F. Frijda, „The emotions", Cambridge University Press, 1986.
[212] N. Harvey, „Wishful thinking impairs belief-desire reasoning: A case of decoupling failure in adults?", Cognition 45 (2), 1992, s. 141–162.

że ludzie oceniają korzyści i straty względem pewnego punktu odniesienia, zamiast absolutnych wartości. Teoria ta odchodzi od tradycyjnej teorii oczekiwanej wartości, która zakłada, że ludzie podejmują decyzje w oparciu o logiczną analizę potencjalnych zysków i strat, a nie o spodziewany wynik. Została opracowana przez psychologów Daniela Kahnemana i Amos Tversky'ego w 1979 roku i jest uważana za jedno z najważniejszych osiągnięć w dziedzinie psychologii ekonomicznej[213].

Zgodnie z teorią perspektywy, ludzie są bardziej wrażliwi na straty niż na korzyści, co oznacza, że bardziej boją się utraty niż cieszą się z zysku. Jest to znane jako efekt niechęci do straty (*loss aversion*). Ponadto ludzie są bardziej skłonni do ryzyka w sytuacji, w której mogą zyskać, niż w sytuacji, w której mogą stracić.

Eksperymenty prowadzone w ramach badań nad teorią perspektywy pozwoliły nazwać trzy inne zjawiska – efekt pewności, efekt odbicia i efekt izolacji. Efekt pewności (*certainty effect*) polega na tym, że ludzie wolą opcje, które gwarantują pewny zysk, nawet jeśli alternatywna opcja oferuje wyższy oczekiwany zysk, ale jest mniej pewna. Efekt odbicia (*reflection effect*) z kolei pokazuje, że ludzie inaczej postrzegają straty i zyski. W przypadku strat ludzie stają się bardziej skłonni do ryzyka, wybierając opcje o niższym oczekiwanym wyniku, ale większej szansie na uniknięcie straty. Natomiast w przypadku zysków ludzie stają się bardziej konserwatywni, preferując opcje o pewnych, choć niższych zyskach. Efekt izolacji (*isolation effect*) polega na tym, że ludzie skupiają się na różnicach między opcjami, a nie na ich podobieństwach, gdy dokonują wyboru. To uproszczenie może prowadzić do niespójnych preferencji i błędnych decyzji. Sposób sformułowania problemu może wpływać na to, które aspekty zostaną uznane za istotne, a co za tym idzie, na wybór danej opcji[214].

2.4 Wnioski

Jak kilkukrotnie zauważono w niniejszym rozdziale, nauki o bezpieczeństwie jako dyscyplina naukowa jest stosunkowo nowym tworem[215]. Temat dezinformacji, manipulacji i wykorzystania przy tym nagrań deepfake, bezsprzecznie wchodzą w jej ramy, natomiast jako zagadnienie było i jest badane również w innych obszarach

[213] D. Kahneman, A. Tversky „Prospect Theory: An Analysis of Decision under Risk", Econometrica XLVII (1979), 1979, s. 263–291.
[214] Tamże.
[215] Powołana została w 2011 roku uchwałą Centralnej Komisji do Spraw Stopni i Tytułów, zmieniającą uchwałę w sprawie określenia dziedzin nauki i dziedzin sztuki oraz dyscyplin naukowych i artystycznych (M.P. 2011 nr 14 poz. 149).

naukowych. Pytanie badawcze postawione w niniejszym rozdziale miało na celu zweryfikowanie obecnego statusu wiedzy dotyczącego fałszywych nagrań deepfake oraz jego wpływu na bezpieczeństwo narodowe. Pytanie badawcze podane na wstępie do niniejszego rozdziału w formie zdania pytającego brzmiało: „jaki jest status ontologiczny wiedzy na temat wpływu zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe?". Próbą odpowiedzi na to pytanie była wysunięta uprzednio hipoteza badawcza, która brzmiała: zmanipulowane materiały audiowizualne mogą mieć istotny wpływ na bezpieczeństwo narodowe, a obecny stan wiedzy na ten temat jest niepełny i nadal rozwijający się. Kompleksowa analiza zgromadzonego w niniejszym rozdziale materiału empirycznego, zebranego przy pomocy metody badawczej jaką jest analiza dokumentów, pozwoliła pozytywnie zweryfikować tę hipotezę.

Przede wszystkim podkreślić należy złożoność i wielowymiarowość problemu jakim jest wpływ zmanipulowanych materiałów audiowizualnych (deepfake'ów) na bezpieczeństwo narodowe. Brakuje ostatecznej zgody co do ontologicznego statusu wiedzy na ten temat, a wiele badań nawzajem sobie przeczy. Jak wykazano, istnieje wiele różnych perspektyw na to zagadnienie, a badania są wciąż w toku.

Zagadnienie deepfake bezsprzecznie niesie ze sobą nowe, potencjalne zagrożenia. Filmy deepfake mogą zostać wykorzystane do rozpowszechniania dezinformacji i propagandy, co może podważyć zaufanie do instytucji i zdestabilizować społeczeństwa. Mogą być również używane do manipulowania opinią publiczną i wpływania na wyniki wyborów. Coraz częściej zdarza się, iż są one wykorzystywane do podszywania się pod osoby publiczne w celu skompromitowania ich reputacji lub próby oszustwa. Przeprowadzono również studium przypadków, gdzie nagrania deepfake używane były do tworzenia fałszywych dowodów w sprawach karnych i cywilnych.

Badacze zidentyfikowali również inne, ważkie zjawisko na pograniczu nauk politycznych i psychologii polityki. Otóż jednostki o poglądach umiarkowanych reagują na fałszywe informacje odmiennie niż ekstremiści. Zjawisko to określono mianem asymetrii ideologicznej[216]. Hipotezę tę potwierdzają badania eksperymentalne, podczas których usiłowano zweryfikować wpływ postaw światopoglądowych na przekazy

[216] L. A. Adamic, N. Glance, „The Political Blogosphere and the 2004 U.S. Election: Divided They Blog", LinkKDD '05 Proceedings of the 3rd International Workshop on Link Discovery, 2005, s. 36-43.

medialne korygujące błędne poglądy[217]. Okazało się, iż osoby o nastawieniu liberalnym skłonne są do korekty poglądów dotyczących posiadania przez Irak broni masowego rażenia, natomiast osoby o postawach centrowych i prawicowych w mniejszym stopniu korygują swoje poglądy pod wypływem nawet uwiarygodnionych faktów[218]. W efekcie obserwatorzy wydarzeń politycznych różnią się w zakresie postrzegania i interpretacji tego samego zbioru faktów.

Badane są również radykalne skutki dezinformacji, w tym potencjał wywoływania konfliktu i generowania przemocy. Planowe obalanie fałszu łagodzi napięcia i sprzyja konstruktywnemu dialogowi. Istotną rolę odgrywają tu narracja: ramowania (*framing*), moralnego wycofania się (*moral disengagement*), pluralistyczna ignorancja (*pluralistic ignorance*). Badacze wskazują na istotną rolę mediów w dostarczaniu informacji, wpływu na sceptycyzm jednostki oraz kluczową rolę jej światopoglądu (bądź uprzednich postaw, trudnych do zmiany)[219]. Należy podkreślić, że pozostaje niejasne, w jakim stopniu, a nawet czy w ogóle, dezinformacja zaostrza polityczną polaryzację. W literaturze akademickiej toczy się na ten temat dyskusja[220], [221], [222]. Podobnie dyskutuje się na temat siły i zakresu oddziaływania na postawy polityczne przez media społecznościowe. Niejasne pozostaje, w jakim stopniu informacja podawana z użyciem Internetu, a w szczególności mediów społecznościowych ma wpływ na postawy polityczne konsumentów tej informacji. Wyniki badań pozostają sprzeczne[223],[224].

Z całą pewnością problem dezinformacji, a zwłaszcza fałszywych nagrań stanowi aktualne i wymagające wyzwanie badawcze. Niezwykle trudno jest ocenić rzeczywisty

[217] D. M. Kahan, „Ideology, Motivated Reasoning, and Cognitive Reflection", Judgment and Decision Making 8, no. 4, 2013, s. 407–424.
[218] B. J. Gaines, J. H. Kuklinski, P. J. Quirk, B. Peyton, J. Verkuilen, „Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq", The Journal of Politics 69 (4), 2007, s. 957-974.
[219] G. Pennycook, T. D. Cannon, D. G. Rand, „Prior Exposure Increases Perceived Accuracy of Fake News", Journal of Experimental Psychology: General 147, no. 12, 2018, s. 1865–1880.
[220] E. Bakshy, S. Messing, L. A. Adamic, „Exposure to Ideologically Diverse News and Opinion on Facebook", Science 348 (6239), 2015, s. 1130-1132.
[221] L. Boxell, M. Gentzkow, J. M. Shapiro, „Greater Internet use Is Not Associated with Faster Growth in Political Polarization among US Demographic Groups", Proceedings of the National Academy of Sciences 114 (40), 2017, s. 10612–10617.
[222] S. Flaxman, S. Goel, J. M. Rao, „Filter Bubbles, Echo Chambers, and Online News Consumption", Public Opinion Quarterly 80 (1), 2016, s. 298-320.
[223] H. Allcott, M. Gentzkow, „Social media and fake news in the 2016 election", Journal of Economic Perspectives 31(2), 2017, s. 1–28.
[224] A. Guess, B. Lyons, B. Nyhan, J. Reifler, „Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Congenial Political News is Less Prevalent than You Think", Knight Foundation report, 2017.

wpływ nagrań deepfake, czy to na jednostkę czy holistycznie na bezpieczeństwo narodowe. Brakuje danych dotyczących skali i zasięgu wykorzystania filmów deepfake w celach zmierzyć zarówno oszustwa jak i dezinformacji. Z każdą chwilą technologia ta staje się coraz bardziej doskonała i powszechnie dostępna. Wraz z każdą nową serią kart graficznych coraz trudniej jest odróżnić prawdziwe materiały audiowizualne od nagrań deepfake. Już teraz ma to duży wpływ na utrudnione rozpoznawanie i zwalczanie dezinformacji.

Jak wykazano w rozdziale, badania nad tak interdyscyplinarnym zjawiskiem wymagają współpracy naukowców z wielu obszarów. Do holistycznej analizy problemu niezbędna jest wiedza nie tylko z dyscypliny nauk o bezpieczeństwie, ale również z obszarów informatyki, psychologii, politologii czy prawa. W obszarze informatyki prowadzone są badania nad metodami automatycznego wykrywania filmów deepfake. Tworzy się również narzędzia mające na celu wspierać użytkownika w procesie rozpoznawania fałszu. Obszar psychologii od lat identyfikuje i nazywa zachowania ludzkie, mające kluczowe znaczenie dla jakości percepcji nagrań przez osoby je oglądające oraz efektów ich oddziaływania na internautów. W zakresie nauk o bezpieczeństwie opracowywane są strategie przeciwdziałania dezinformacji i propagandzie rozprzestrzenianej za pomocą filmów deepfake. Natomiast obszar prawny bada implikacje prawne i etyczne związane z wykorzystaniem nagrań deepfake, tworząc ramy prawne mające na celu ograniczenie negatywnych skutków tej technologii. Jednak dopiero połącznie wiedzy i doświadczenia każdej ze wskazanych dyscyplin pozwala na pełne zrozumienie problemu i prowadzenie nad nim kompleksowych badań.

Wiedza na temat wpływu nagrań deepfake na bezpieczeństwo narodowe jest wciąż w fazie rozwoju. Istnieje wiele potencjalnych zagrożeń związanych z wykorzystaniem nagrań deepfake i ciężko jest je wszystkie zidentyfikować. Konieczne są dalsze badania, aby lepiej zrozumieć to zjawisko i opracować skuteczne strategie przeciwdziałania mu.

Rozdział 3. Techniczne aspekty nagrań deepfake oraz rozwój technologii

Do przeprowadzenia badań nad zagrożeniem nagraniami deepfake dla bezpieczeństwa narodowego, konieczne było wpierw przebadanie wykorzystanej technologii w celu wykrycia jej możliwości, słabych i mocnych stron oraz do analizy potencjału dalszego rozwoju. Dzięki zapoznaniu się z możliwościami dostępnych narzędzi, wybrano najbardziej profesjonalne oprogramowanie i na jego podstawie przygotowano nagrania wykorzystane w eksperymencie. Niniejszy rozdział poświęcony został analizie możliwości tworzenia nagrań deepfake i prezentuje szczegółowe informacje umożliwiające zrozumienie wykorzystanego narzędzia.

Czyniąc zadość wymogom metodologicznym stawianym pracom badawczym w dyscyplinie nauk o bezpieczeństwie, sformułowano właściwy dla niniejszego rozdziału problem badawczy. Zgodnie z przyjętą w pracy metodologią, problem został sformułowany w postaci zdania pytającego i brzmi następująco: w jaki sposób powstają nagrania deepfake? Ma on na celu nie tylko ukazanie, jak przebiega powstawanie materiałów, lecz również to, w jaki sposób powstały materiały wykorzystane w badaniu (2 nagrania, wykorzystujące wizerunki dwoje znanych influencerów).

Dopełnieniem prezentowanego problemu szczegółowego jest niniejsza hipoteza: nagrania deepfake powstają poprzez wykorzystanie oprogramowania do zmiany obrazu lub dźwięku w oryginalnym nagraniu, tak aby wyglądało to jakby ktoś inny mówił lub wyglądał inaczej niż w rzeczywistości. Weryfikacja hipotezy zawiera się w niniejszym rozdziale, a jej badanie ma bezpośredni wpływ na główny problem badawczy, prezentowany w niniejszej dysertacji.

3.1 Przygotowanie stanowiska roboczego

Na potrzeby przygotowania niniejszej pracy oraz odpowiedzi na sformułowane szczegółowe pytanie badawcze i weryfikacji stawianej hipotezy, konieczne było własnoręczne przygotowanie nagrań deepfake, a przez to wpierw prześledzenie i opisanie całego procesu.

Zastanawiające dla autora było, czy przy ówczesnych możliwościach sprzętowych (2020 rok), przeciętny użytkownik komputera jest w stanie wygenerować dobrej jakości nagrania, trudne lub niemożliwe do odróżnienia oraz w jaki sposób proces

tworzenia przebiega. Nie było więc celem utworzenie idealnych nagrań z pomocą wyspecjalizowanego sprzętu, a jedynie średnio prezentujące się filmy, powstał z wykorzystaniem pojedynczej, komercyjnej karty graficznej, wykorzystywanej powszechnie w grach komputer.

Celem weryfikacji hipotezy, postanowiono utworzyć w warunkach laboratoryjnych średniej klasy stanowisko gamingowe i przy jego pomocy utworzyć nagrania testowe. Na realizację tej potrzeby niezbędny był zakup i złożenie średniej klasy komputera, którego cena po oszacowaniu części miała wynosić do 10 000 zł. Wielkość kwoty wynika przede wszystkim z wysokich w owym czasie kosztów kart graficznych, spowodowanych pandemią i ograniczeniami w produkcji podzespołów.

Na stanowisku zainstalowano kilka rodzajów oprogramowania służącego do tworzenia nagrań deepfake, a następnie przetestowano każde z nich. Ostatecznie wybrano jedną aplikację i przystąpiono do tworzenia fałszywych nagrań.

3.1.1 Organizacja laboratorium badawczego

Na potrzeby niniejszych badań zdecydowano się złożyć stacjonarny komputer, który swoimi parametrami pozwalałby na swobodne tworzenie nagrań deepfake, jednak z charakteru przypominał standardowe wyposażenie współczesnego gracza. Każdy z elementów dobierany był indywidualnie, tak by całość oferowała jak najlepsze efekty przy jak najmniejszym koszcie.

Podstawą i najważniejszym elementem przy tworzeniu nagrań deepfake jest odpowiednio silna karta graficzna. Niewystarczająca ilość pamięci karty graficznej (VRAM) powoduje ograniczenia jakościowe w tworzonych materiałach[225]. Kolejnym ogranicznikiem wyboru karty była jej marka. W tworzeniu nagrań deepfake najlepiej sprawdzają się karty marki GeForce, które zapewniają pełne wsparcie dla technologii CUDA.

Nagrania deepfake tworzone przez profesjonalistów, najczęściej powstają na kartach graficznych wykorzystywanych na co dzień do pracy ze sztuczną inteligencją, w symulacjach i obliczeniach na dużą skalę (zwłaszcza obliczeniach zmiennoprzecinkowych) oraz w zaawansowanym generowaniu obrazów w dziedzinach

[225] A. Dodge, „Using Fake Video Technology To Perpetuate Intimate Partner Abuse – Domestic Violence Advisory”, https://www.cpedv.org/sites/main/files/webform/deepfake_domestic_violence_advisory.pdf [dostęp: 01.10.2022].

zawodowych i naukowych. Obecnie, ze względu na duży rozmiar pamięci graficznej, jedną z lepszych do tworzenia fałszywych nagrań jest karta NVIDIA A100 oferująca 80GB VRAM[226]. Koszt takiej karty to ok. 70 000 zł, co jest kwotą zaporową dla indywidualnych użytkowników. Ponadto karty serii Tesla/Datacenter nie nadają się do indywidualnego użycia. Ich specyficzna budowa nie pozwala na wykorzystanie jej w domowych warunkach do grania w nowsze gry czy oglądania wysokiej jakości filmów. Pozwala ona za to na trenowanie modeli w wyższej rozdzielczości i przy zastosowaniu lepszych ustawień. Czas trenowania jest też nieporównywalnie krótszy od czasu trenowania na pojedynczych kartach dostępnych standardowo w sklepie.

Często mniej profesjonalni twórcy, tworząc swoje stanowiska pracy, wykorzystują do ich budowy kilka kart graficznych, łącząc je ze sobą równolegle. Popularne jest na przykład łączenie dwóch lub trzech kart NVIDIA 3090, dzięki czemu można otrzymać średniej prędkości zestaw oferujący aż 72 GB pamięci VRAM[227].

Korzystanie z Deep Face Lab 2.0 wymaga komputera o wysokiej wydajności z nowoczesnym procesorem graficznym, obszerną pamięcią RAM, pamięcią masową i szybkim procesorem. Windows 10 jest ogólnie zalecany dla większości użytkowników, ale bardziej zaawansowani użytkownicy mogą chcieć używać Linuksa, aby uzyskać lepszą wydajność. Przy obecnych kartach graficznych jest ona w praktyce niezauważalnie większa, gdyż obciążenie interfejsu Windowsa w niewielkim stopniu obciąża kartę. Pomysłem na optymalizację może być uruchamianie systemu na odrębnej, np. zintegrowanej karcie graficznej, zaś przeznaczenie całości mocy drugiej karty na obliczenia.

Aby uzyskać optymalną wydajność, procesor powinien posiadać co najmniej 4 rdzenie (zalecane jest 8 lub więcej) minimum 6 generacji, 16 GB pamięci RAM, szybką, dedykowaną pamięć SSD (500 GB-1 TB). Do przechowywania aktualnie przetwarzanych plików i modeli/zestawów danych. Zalecane jest również upewnienie się, że systemowy rozmiar pliku stronicowania jest wystarczająco duży, aby zapobiec problemom z zapisywaniem modeli i ogólną wydajnością (minimum 32 GB, 64 GB lub nawet 128 GB, jeśli używa się 2-3 GPU w jednym systemie, najlepiej również na SSD). Proces szkolenia wymaga długiego działania komputera, zalecane jest więc również dobre

[226] Karta ta jest wykorzystywana między innymi przez autorów (startup) kanału deeptomcruise, tworzącego profesjonalne nagrania z wykorzystaniem wizerunku Toma Cruise. Nagrania dostępne między innymi na stronie: https://www.tiktok.com/@deeptomcruise [dostęp: 12.2022].
[227] „GPUs Multiple with Learning Deep”, [dostęp: 10.09.2022].

chłodzenie i wysokiej jakości zasilacz ze sporym zapasem mocy. W przypadku chłodzenia standardowego – powietrzem – powinno unikać się używania DFL przez dłuższy czas. Podobna reguła dotyczy urządzeń przenośnych – używanie DFL na laptopie spowodować może generowanie nadmiernego ciepła, przez co długotrwałe używanie DFL może skrócić żywotność sprzętu.

Uwzględniając powyższe aspekty, jak również mając na uwadze ograniczenia budżetowe, zdecydowano się na złożenie jednostki z wymienionych poniżej elementów.

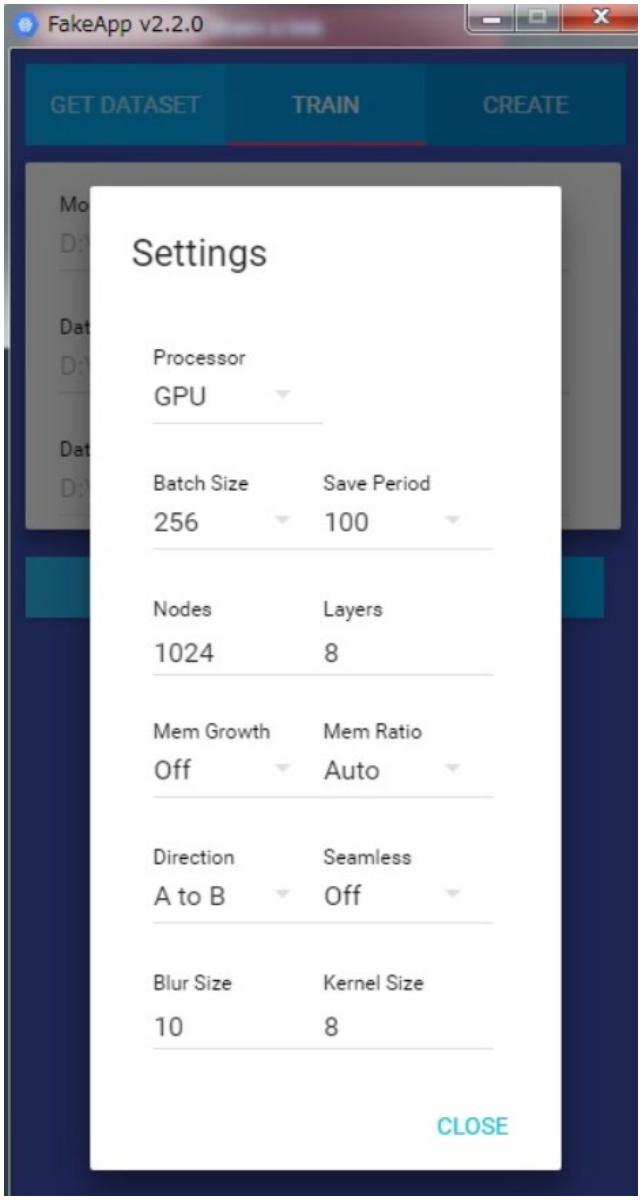
- Płyta główna: MSI MAG B560 TOMAHAWK WIFI
- Procesor: Intel® Core™ i9-11900F Rocket Lake 2.5 GHz/5.2 GHz 16MB LGA1200 BOX
- Pamięć operacyjna: DDR4 G. Skill Ripjaws V 32GB (2x16GB) 3200MHz CL16 1,35V
- Karta graficzna: Asus TUF GeForce RTX 3060 Gaming OC 12GB GDDR6
- Dysk Twardy 1: SSD M2 Silicon Power A80 512GB PCIe Gen3x4 NVMe (3400/3000 MB/s) 2280
- Dysk Twardy 2: SEAGATE SkyHawk™ 2TB ST2000VX008 64MB SATA III
- Obudowa: Fractal Design Define S2 czarna
- Zasilacz: be quiet! Straight Power 11 850W 135mm 80+Gold
- Peryferia: Chłodzenie wodne NZXT Kraken X63 280mm

3.1.2 Przegląd dostępnych aplikacji i technologii

Do tworzenia fałszywych nagrań, manipulacji obrazu i dźwięku, powstało dotychczas wiele aplikacji, zarówno na komputery jak i telefony. Wybór zależy w głównej mierze od posiadanych zasobów, technicznej wiedzy oraz oczekiwanego efektu, a także posiadanego czasu do pracy nad nagraniem. Poniżej opisano powszechnie dostępne aplikacje oraz porównano charakterystykę użycia oraz działanie.

Pierwszą aplikacją, jaka pojawiła się w sieci dla ogółu użytkowników, była aplikacja FakeApp. Opublikowana została w 2017 roku przez użytkownika Reddit-a, ukrytego pod pseudonimem „DeepFakeapp”. Udostępnia interfejs graficzny użytkownika (GUI) do tworzenia nagrań deepfake. Aplikacja jest łatwa w obsłudze, ale przez to

o ograniczonych możliwościach. Wsparcie dla aplikacji zostało wstrzymane i dostępna jest wyłącznie stara, nieoficjalna wersja v2.2.0[228]. Nie zaleca się jej używania.



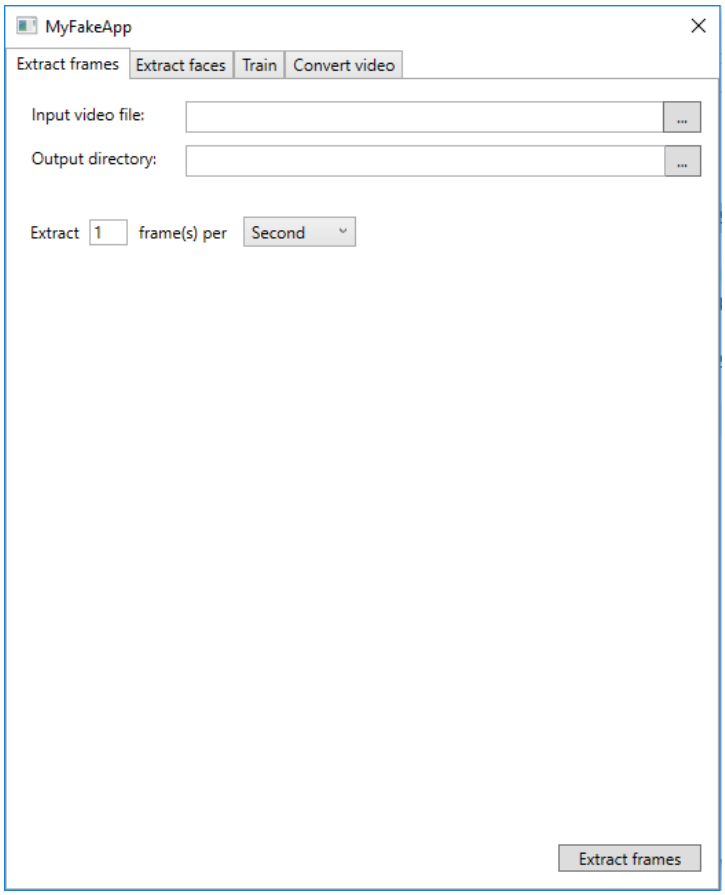
Grafika 1 Interfejs aplikacji FakeApp w ostatniej dostępnej wersji v2.2.0. Opracowanie własne.

Na powyższym screenie widzimy wstępną konfigurację aplikacji – możliwy jest samodzielny wybór elementu do trenowania – karty graficznej lub procesora, a także pozostałe ustawienia dotyczące treningu obrazu.

Kolejną aplikacją, która kilka miesięcy później została powszechnie udostępniona w sieci, była aplikacja MyFakeApp, stworzona przez użytkownika Radek Hlaváček[229].

[228] Tamże.
[229] Blog twórcy oprogramowania MyFakeApp, https://radek350.wordpress.com/2018/02/17/myfakeapp-fakeapp-alternative/ [dostęp: 01.01.2023].

Charakteryzował ją odświeżony interfejs, rozbudowane funkcje, a także poprawiona konfiguracja. Aplikacja doczekała się tylko jednej aktualizacji i po pewnym czasie również została porzucona.



Grafika 2 Interfejs aplikacji MyFakeApp w starszej wersji. Opracowanie własne.

Obie aplikacje wymienione powyżej cechowała prostota użytkowania oraz przyjazny dla niedoświadczonego użytkownika interfejs obsługi. Odbywało się to jednak kosztem ograniczonej funkcjonalności i małą ilością opcji. Obecnie podobny efekt otrzymać można korzystając z aplikacji deepfake dostępnych na telefony lub jako aplikacje web-owe.

Kilka miesięcy później na rynku deepfake pojawiły się dwa, konkurujące ze sobą do dzisiaj, projekty – DeepFaceLab oraz FaceSwap. Obie są zbliżone do siebie pod względem funkcjonalności oraz poziomem zaawansowania. Obie aplikacje mają również podobną liczbę gwiazdek oceniających na github-ie – ponad 40 tysięcy[230]. Obie mają również zbliżoną liczbę rozgałęzień – DeepFaceLab 9 tysięcy, zaś FaceSwap 12

[230] Dla aplikacji DeepFaceLab jest to około 42 tysiące, zaś dla FaceSwap 46 tysięcy, https://github.com/deepfakes/faceswap https://github.com/iperov/DeepFaceLab [dostęp: 01.06.2023].

tysięcy[231]. Wnioskować po tym można, iż przy obu aplikacjach istnieje zbliżona wielkością liczba

Wykorzystaną w eksperymencie aplikacją, jest DeepFaceLab (DFL). Jest to open-source'owy projekt, który udostępnia narzędzia do tworzenia nagrań deepfake. Posiada interfejs linii poleceń i wymaga podstaw wiedzy i umiejętności z zakresu sztucznej inteligencji i uczenia maszynowego. Jest to również najbardziej zaawansowana aplikacja, dostępna dla nieprofesjonalnych użytkowników w licencji open-source. Co więcej cechuje się ona bogatą społecznością – dostępnych jest kilka forum oraz grup na komunikatorach społecznych[232]. Z powyższych względów został on wybrany do tworzenia nagrań zarówno dla eksperymentu, jak i do niniejszej pracy.

Deepfakes Web: jest to przeglądarkowa aplikacja do tworzenia deepfake'ów, która pozwala na przetwarzanie plików wideo bez konieczności pobierania i instalowania specjalistycznego oprogramowania. Aplikacja jest płatna[233].

GANFakes: przez wiele osób narzędzie błędnie nazywane jest tworzącym deepfake-i. W rzeczywistości generowane tam twarze są całkowicie nowymi twarzami, zbudowanymi na podstawie analizy setek tysięcy twarzy innych osób. Aplikacja korzysta z algorytmów GAN (Generative Adversarial Networks), a wytworzone przez nią obrazy pobrać można ze strony twórcy[234]. Projekt zamieszczono w tym miejscu, ze względu na jego łatwe wykorzystanie, a także popularność w tworzeniu oszukańczych profili.

Deepfake-TIMIT – jest to narzędzie do tworzenia deepfake'ów, które zostało specjalnie zaprojektowane do przetwarzania dźwięku. Jest dostępne jako aplikacja na platformie Windows.

Deepfake-o-Matic – to aplikacja do tworzenia deepfake'ów, która jest dostępna jako rozszerzenie do przeglądarki Chrome.

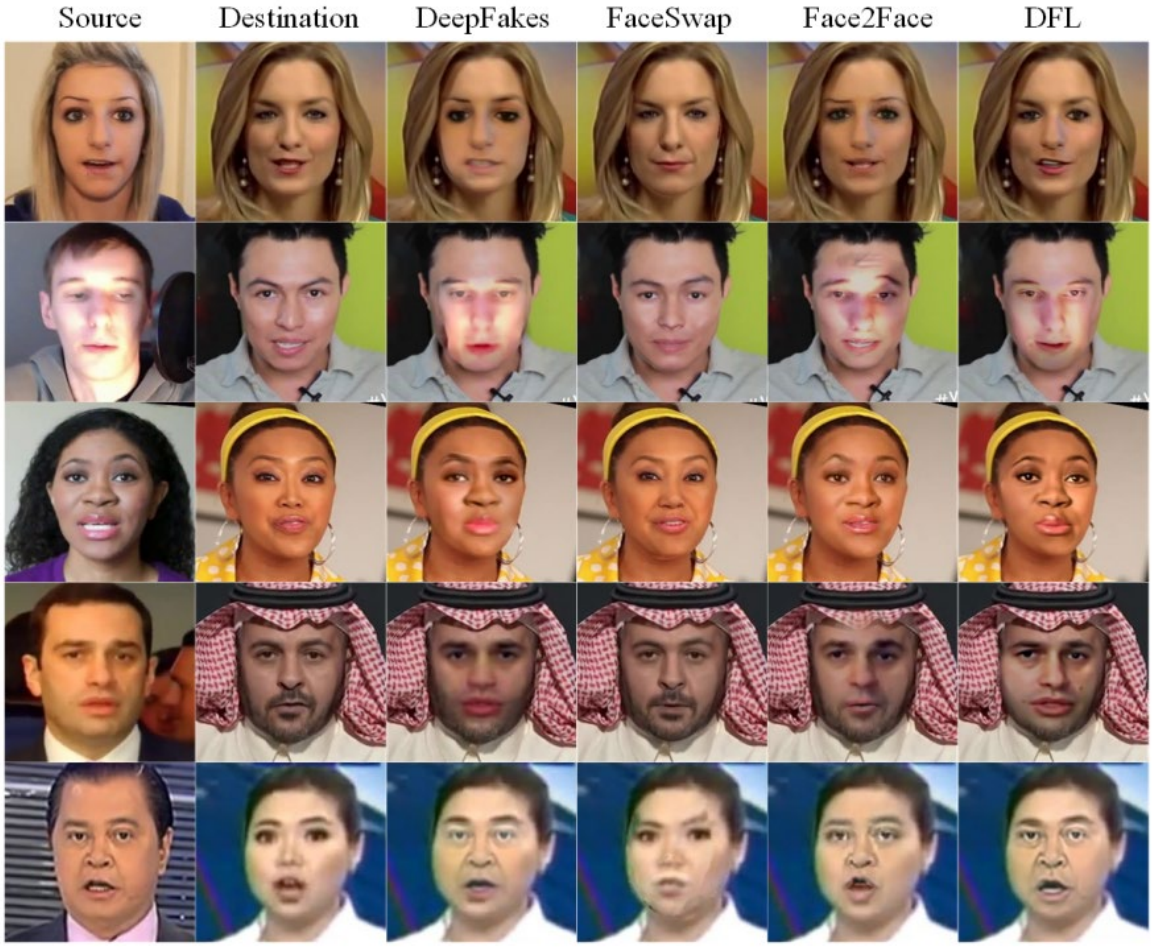
DeepArt – to narzędzie do tworzenia deepfake'ów, które korzysta z algorytmów sieci neuronowych do generowania obrazów.

VideoSwap – jest to aplikacja do tworzenia deepfake'ów, która jest łatwa w obsłudze i pozwala na łatwe zamienianie twarzy w filmach. Aplikacja oferuje również

[231] Tamże.
[232] J. Pu, N. Mangaokar, „Deepfake Videos in the Wild: Analysis and Detection", 2021.
[233] Strona główna aplikacji Deepfakes Web, https://deepfakesweb.com/plan [dostęp: 01.01.2023].
[234] Strona główna projektu firmy Stability.AI, https://thispersondoesnotexist.com/ [dostęp: 01.01.2023].

narzędzia do korekty i poprawy jakości generowanych filmów. Dostępna jest na platformie iOS.

Ze względu na chęć wybrania aplikacji potencjalnie mogącej utworzyć trudne do rozpoznania nagranie, zdecydowano się na przegląd analiz porównawczych wymienionych aplikacji. Poniżej zaprezentowano zestawienie efektów tworzenia nagrań deepfake dla czterech wybranych aplikacji – DeepFakes, FaceSwap, Faces2Faces oraz DeepFaceLab.



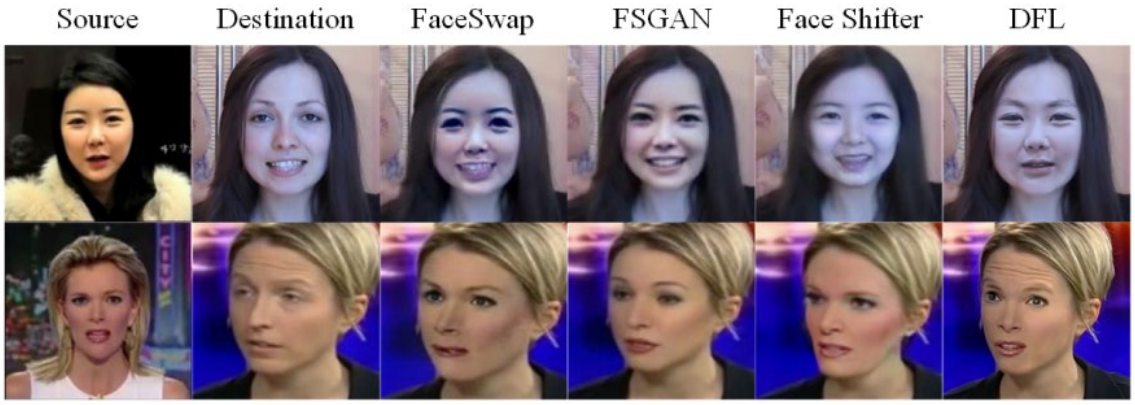
Grafika 3 Zestawienie twarzy wygenerowanych przy pomocy aplikacji deepfake typu open – source[235].

Oglądając utworzone kadry nagrań, zauważyć można, iż obrazy utworzone przy pomocy DeepFaceLab w największym stopniu przypominają cel. Niewiele słabszy okazała się aplikacja FaceSwap, oferująca zbliżony poziom zaawansowania.

[235] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, J. Jiang, L. RP, S. Zhang, P. Wu, W. Zhang, „DeepFaceLab: A simple, flexible and extensible face swapping framework", 2020.

Grafika 4 Zestawienie twarzy wygenerowanych przy pomocy aplikacji deepfake – mężczyzna[236].

Gorszej jakości zdają się być obrazy wytworzone przy pomocy pozostałych dwóch aplikacji – DeepFakes oraz Face2Face. Zwłaszcza to pierwsze narzędzie, ze względu na brak wsparcia twórcy oraz wiekowość projektu, zdaje się nie być zdolnym do oszukania kogokolwiek.



Grafika 5 Zestawienie twarzy wygenerowanych przy pomocy aplikacji deepfake – kobieta[237].

3.1.3 Wybór i instalacja wybranego oprogramowania

DeepFaceLab to open-source'owy system deepfake stworzony przez {iperov} do zamiany twarzy z ponad 3000 rozwidleniami i 13000 gwiazdami na Github: zapewnia niezbędną i łatwą w użyciu linię dla ludzi bez pełnego zrozumienia głębokiego framework do nauki lub z implementacją modelu, a jednocześnie pozostaje elastyczną i luźną strukturą łączącą dla osób, które muszą wzmocnić swój potok innymi funkcjami bez pisania skomplikowanego kodu wzorcowego.

[236] Tamże.
[237] Tamże.

Aplikacja dostępna jest na kanale Github autora[238] i możliwa do pobrania bezpośrednio z serwisu, lub dysków chmurowych, takich jak Mega.nz czy dysk Yandex. Aplikacja dostępna jest zarówno na systemy Windows jak i Linux, a także możliwe jest skorzystanie z gotowego repozytorium dla Google Colab. Aplikacja na dzień 01.01.2023 posiadała 38 tysięcy gwiazdek oraz prawie 9 tysięcy kopii repozytorium. Autor na profilu, chwali się, że przy pomocy jego aplikacji powstaje aż 95% wszystkich nagrań deepfake, jednak informacja ta jest trudna do zweryfikowania.

Celem instalacji aplikacji na systemie Windows, niezbędne jest pobranie najnowszej wersji aplikacji (aktualne linki znajdują się na profilu autora), a następnie rozpakowanie archiwum. W dalszych krokach należy upewnić się, że posiadamy zainstalowane najnowsze sterowniki karty graficznej oraz CUDA w wersji minimum 10.

3.2 Tworzenie nagrań deepfake – obraz.

Powstało wiele aplikacji i skryptów do tworzenia nagrań deepfake. Najpopularniejsze, oferowane obecnie na rynku to Face Swap oraz DeepFakes. Jednak jak opisano w powyższych podrozdziałach, aplikacją z największą ilością możliwości, najbardziej zaawansowaną w rozwoju, jest DeepFaceLab. Program dostępny jest na licencji open source i rozwijany przez użytkowników. Pozwala swobodnie mieszać różne mechanizmy technologii deepfake, na każdym etapie tworzenia wideo.

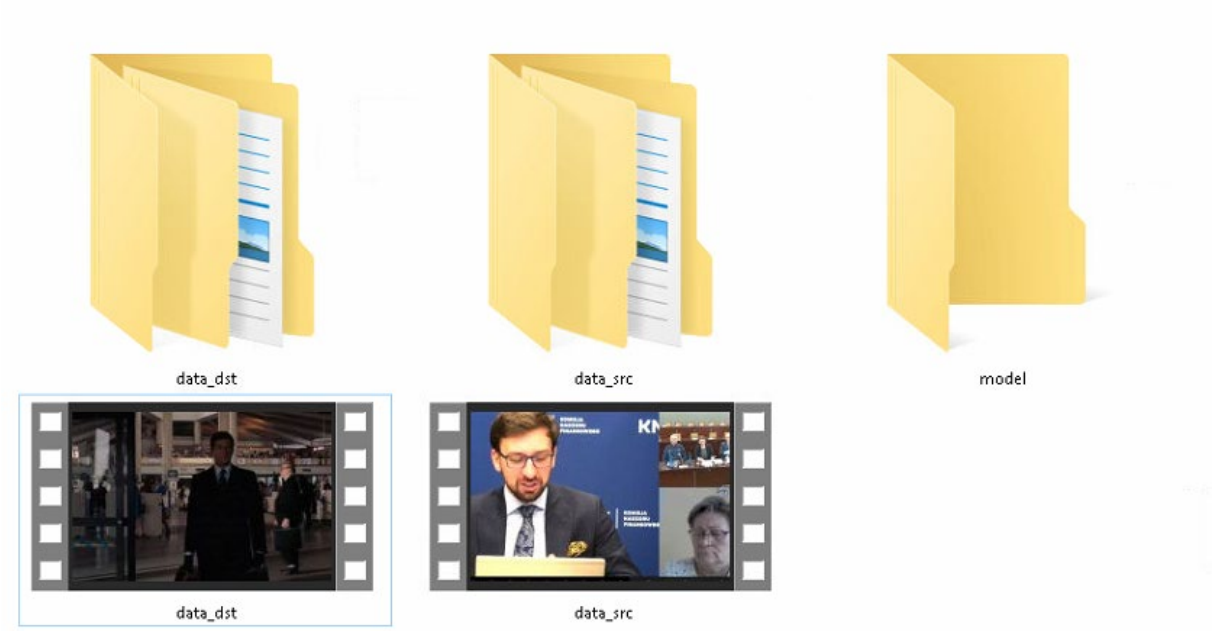
Tworzenie nagrania w przypadku każdej z wymienionych powyżej aplikacji, składa się z podobnych 7 etapów i konieczne jest wykonywanie ich po kolei, bez pominięcia żadnego. Ze względu na brak polskiej terminologii oraz celem uszczegółowienia pojęć, poszczególne nazwy etapów zaprezentowane są przy pomocy anglojęzycznych nazw, które podano w nawiasach.

3.2.1 Przygotowanie miejsca pracy

W folderze workspace konieczne jest przygotowanie przestrzeni roboczej dla plików. Należy utworzyć dwa, puste foldery o nazwach data_scr (folder danych źródłowych) oraz data_dst (folder danych docelowych). Będą one miejscem przechowywania klatek wyciągniętych z wideo źródłowego (data_src.mp4 – wideo, z którego wyodrębnia się twarz, którą będzie nakładana na wideo docelowe) oraz wideo

[238] Strona Github autora projektu Ivana Petrova, https://github.com/iperov/DeepFaceLab [dostęp: 01.01.2023].

docelowego (data_dst.mp4 – wideo, na które zostanie nałożony obraz źródłowy). W przypadku gdy robimy kolejne wideo, pomocnym może być polecenie „clear workspace”, które automatycznie wyczyści używaną przestrzeń i przystosuje ją pod kolejny projekt.



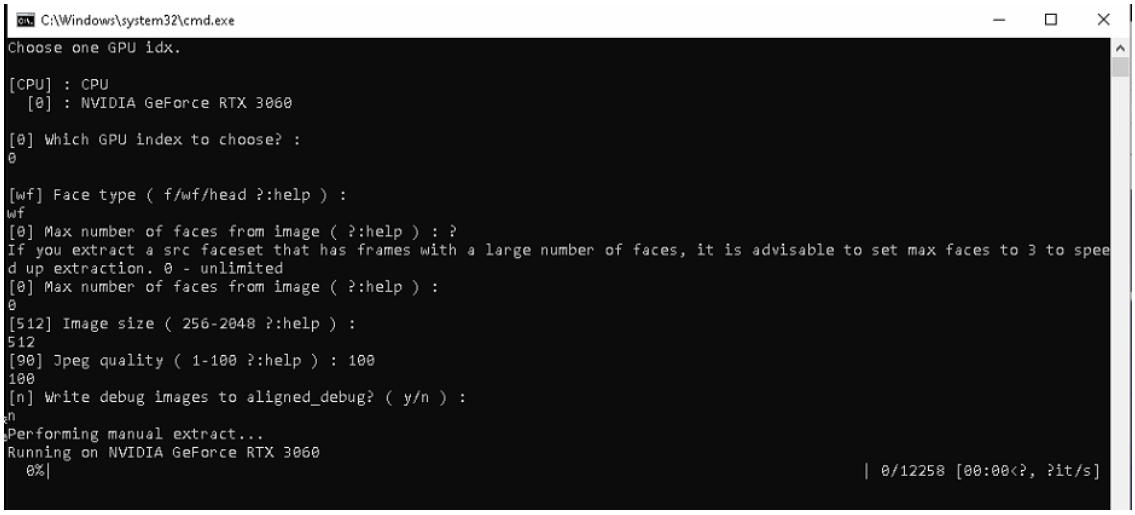
Grafika 6 Uporządkowane miejsce pracy – foldery z zawartością. Opracowanie własne.

3.2.2 Wyodrębnianie klatek z wideo

Wyodrębnianie obrazów z wideo źródłowego – data_src – to pierwszy krok przy robieniu nagrań deepfake. Pomocnym może być przy tym opcjonalna funkcja aplikacji *cut video,* która pozwala szybko przyciąć dowolny film do żądanej długości, upuszczając go na ten plik .bat. Przydatna zwłaszcza wówczas, gdy nie posiada się żadnego oprogramowania do edycji wideo.

Wyodrębnione klatki z wideo data_src.mp4 program automatycznie umieszcza w folderze „data_src”. Dostępne są 2 formaty docelowe klatek: jpg oraz png. Zalecane jest stosowanie plików jpg, które zajmują mniej miejsca, a ich jakość jest wystarczająco dobra. Pliki z rozszerzeniem png są dużo większe, a nie oferują znacznie wyższej jakości. Być może w przyszłości konieczne będzie stosowanie formatu png, jednak ze względu na ograniczenia sprzętowe, na potrzeby niniejszego badania wybrano format jpg. Dla porównania rozmiar folderu zawierającego obrazy w formacie png zajmuje do 8 razy więcej miejsca niż ten sam folder zawierający klatki w formacie jpg. Czas trwania procesu zadania uzależniony jest od długości nagrania, jego oczekiwanej jakości, zgromadzonych materiałów oraz mocy procesora.

Podobnie jak w przypadku wideo źródłowego i tym razem program oferuje możliwość wyodrębnienia obrazów z wideo data_dst. Klatki z pliku data_dst.mp4 umieszczane są w folderze data_dst, w wybranym przez nas formacie. Do wyboru są jpg/png i podobnie jak w przypadku wideo źródłowego, zalecane jest wybranie rozszerzenia jpg.



Grafika 7 Wyodrębnianie twarzy z klatek filmu. Opracowanie własne.

3.2.3 Wyodrębnianie masek twarzy z klatek wideo źródłowego

Pierwszym etapem przygotowania źródłowego zestawu danych jest wyodrębnienie twarzy z wyodrębnionych ramek znajdujących się w folderze „data_src". Jest to niezbędne, by prawidłowo przeprowadzić dalszy proces treningu modelu.

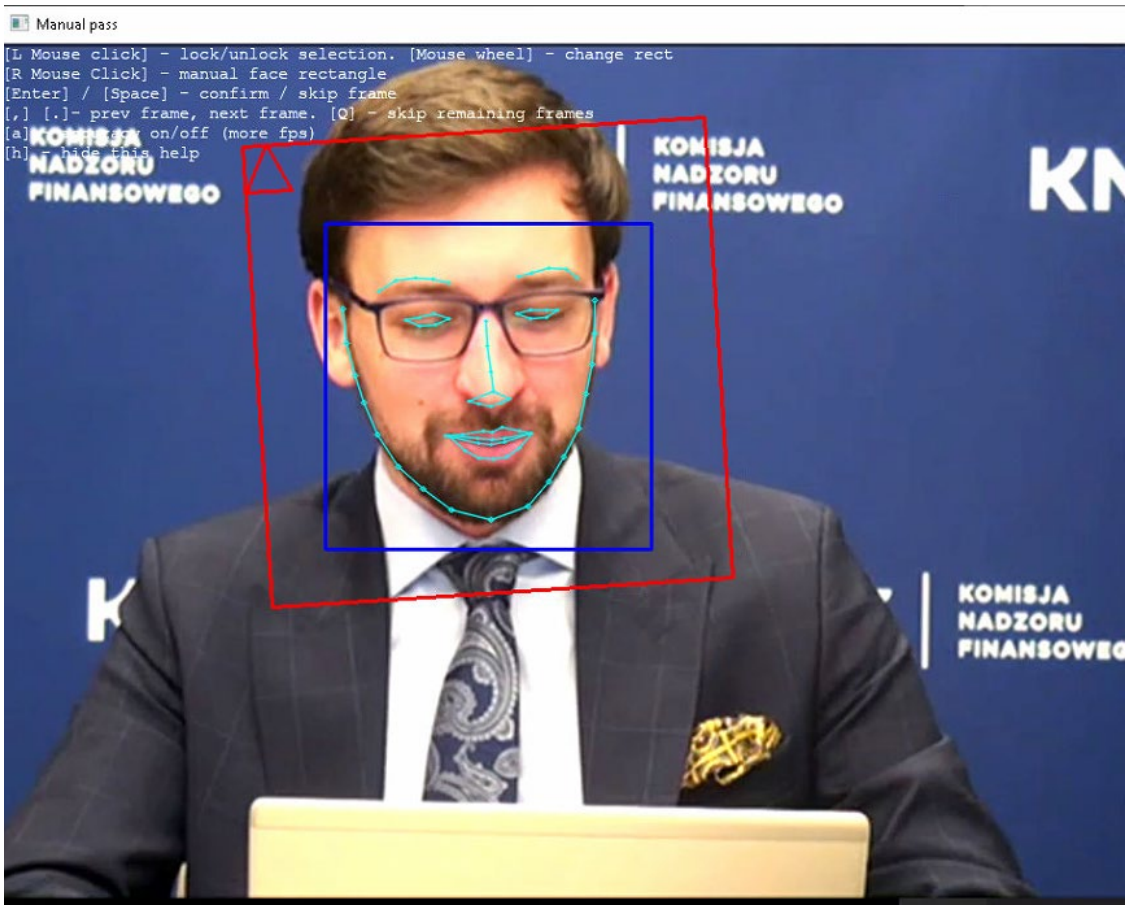
Dostępne są 3 techniki wyodrębniania twarzy, a wybór prawidłowej zależny jest od oczekiwań twórcy, jego umiejętności oraz możliwości sprzętowych. Najmniej wymagającym jest wyodrębnienie pełnej twarzy (*full face* – FF). W zależności od możliwości materiału źródłowego, FF należy wybrać dla całej twarzy (HF) jak i połowy twarzy (MF), nie zaleca się używania go w zestawieniu z całą twarzą FF.

Drugą z możliwości jest wyodrębnienie całej twarzy (whole face – WF). Model ten zawierać będzie wówczas więcej elementów, takich jak podbródek czy kości policzkowe i zalecany jest dla modeli całej twarzy (WF) lub niższych. Jest to uniwersalne i najczęściej wybierane rozwiązanie do pracy zarówno z modelami FF jak i WF.

Najbardziej zaawansowane jest wyodrębnianie całej głowy (head – HEAD), gdzie otrzymujemy również uszy i włosy modela. Obecnie HEAD zalecana jest do treningu jedynie zaawansowanym użytkownikom, gdyż w trakcie treningu wystąpić może wiele

komplikacji z prawidłowym odzwierciedleniem włosów. Maski HEAD należy trenować wyłącznie z maskami HEAD, nie należy stosować tu innych rozwiązań.

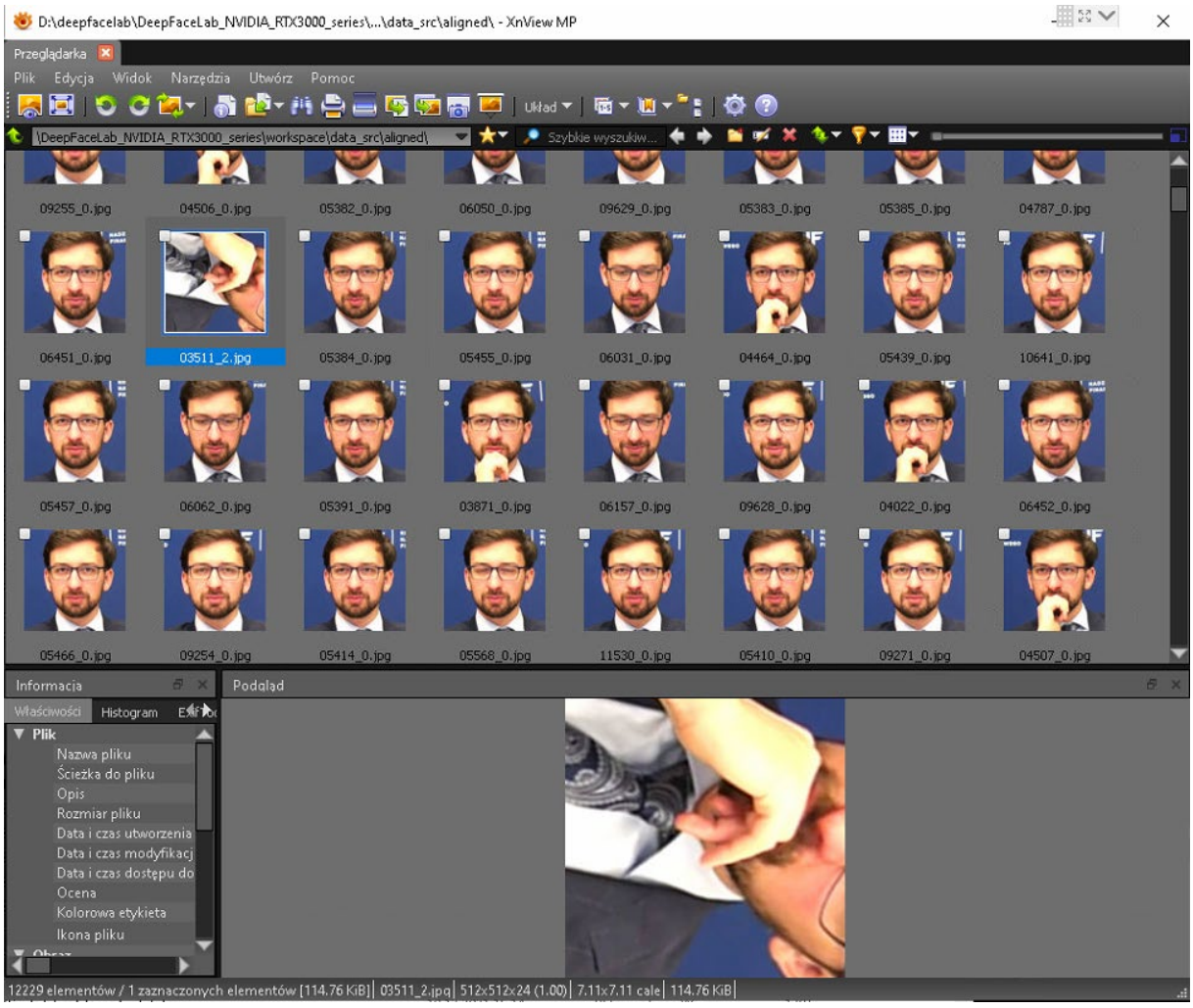
Dostępne są dwa tryby wyodrębniania twarzy – automatyczny i manualny. Manualny rekomendowany jest w przypadku dużej ilości twarzy na danym obrazie jak i ich słabej jakości. Pozwala on ręcznie wybrać twarz mającą podlegać dalszej obróbce jak i oznaczyć szczegółowo elementy wchodzące w jej skład.



Grafika 8 Rozpoznawanie twarzy na klatkach filmu. Opracowanie własne.

Oczyszczanie klatek wideo źródłowego

Po zakończeniu wyodrębniania masek twarzy, następnym krokiem jest wyczyszczenie źródłowego zestawu twarzy/zestawu danych z fałszywych trafień/niepoprawnie wyrównanych twarzy. DeepFaceLab oferuje przy tym kilka przydatnych narzędzi. Jednym z nich jest opcja „data_src view aligned result", która otwiera zewnętrzną aplikację, a ta pozwala szybko przejrzeć zawartość folderu „data_src/aligned". Znajdują się w niej wszystkie wyodrębnione twarze. Należy przejrzeć je pod kątem pomyłek systemu i niepoprawnie wyrównanych twarzy źródłowych. Należy również wyszukać twarze innych osób, tak by nie zanieczyszczały treningu maski. Niepoprawnie rozpoznane twarze należy poprawić lub usunąć.



Grafika 9 Poklatkowa weryfikacja zidentyfikowanych twarzy. Opracowanie własne.

Narzędzie „data_src sort" zawiera wbudowane różne algorytmy sortowania, które pomocne są w znajdywaniu niechcianych elementów. Dostępne są opcje filtrowania twarzy na podstawie ich nietypowości. Pojedyncza twarz, przypadkowo wyodrębniona, wyróżniać się bowiem może od głównej twarzy wieloma aspektami, takimi jak jakość obrazu, nasycenie barwą czy po prostu rozmiarem. Narzędzie „data_src sort" pozwala filtrować obrazy między innymi na podstawie rozmycia ruchu, kierunku odchylenia twarzy, kierunku nachylenia twarzy, rozmiarach prostokąta, w którym twarz jest umieszczona, podobieństwie histogramu, niepodobieństwie histogramu, jasność obrazu, odcienia, ilości czarnych pikseli, oryginalnej nazwie pliku, ilości twarzy na oryginalnej klatce czy bezwzględnej różnicy pikseli.

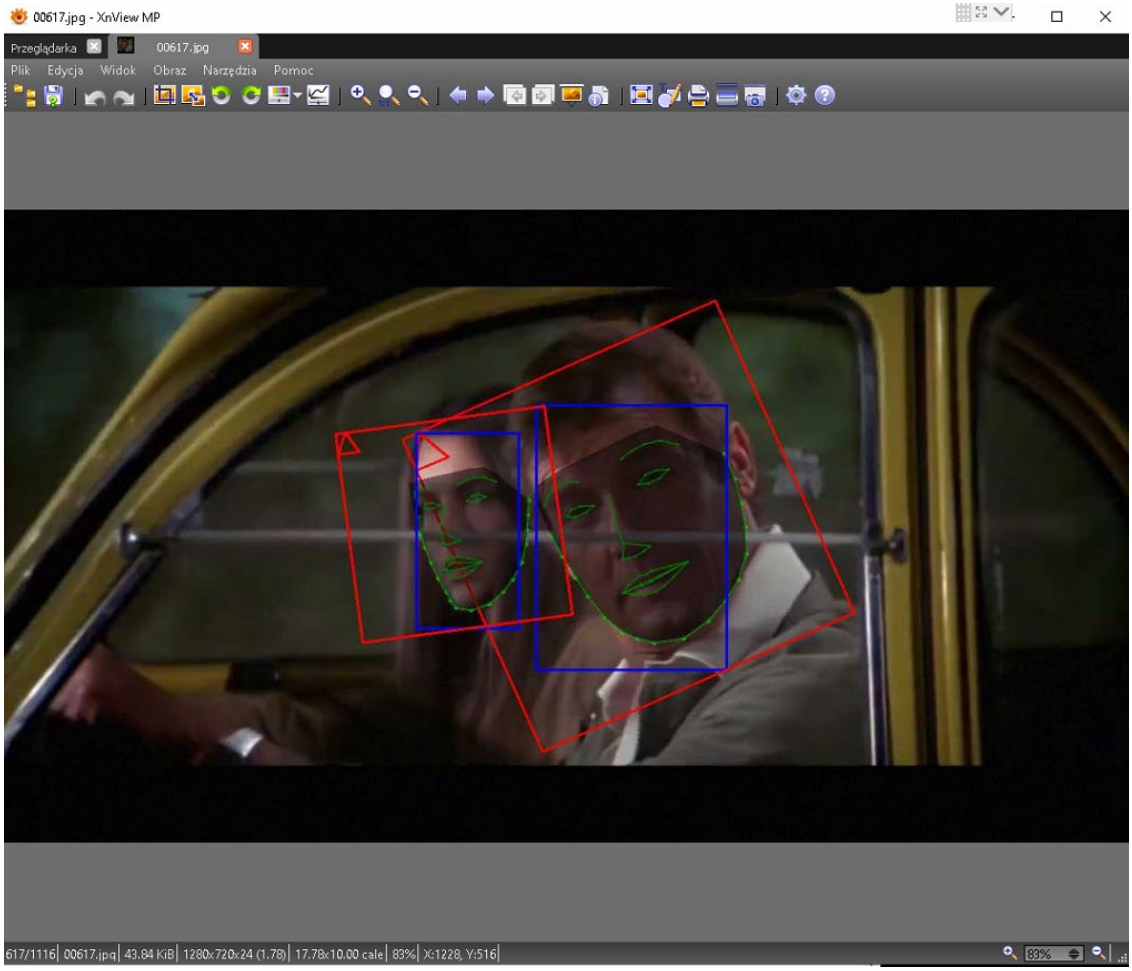
Narzędzie „data_src util add landmarks debug images" pozwala wygenerować folder „aligned debug", po prawidłowym wyodrębnieniu wszystkich twarzy. Uzyskanie tych elementów możliwe jest również w trakcie wyodrębniania twarzy, jednak dla data_src jest on domyślnie wyłączony. „aligned debug" wzmocni proces dostosowywania

danych źródłowych do obrazu docelowego w celu stworzenia bardziej realistycznego deepfake.

Narzędzie „data_src util faceset enhance” pozwala na użycie specjalnego algorytmu uczenia maszynowego do tworzenia ekskluzywnych lub ulepszonych (upscale/enhance) wyglądów twarzy w danym zestawie danych. Jest to przydatne zwłaszcza, jeżeli zestaw danych jest nieco rozmazany lub niepełny lub chce się, by docelowy obraz miał jeszcze więcej szczegółów/tekstur. Zarówno „faceset enhance” jak i ”aligned debug” nie są niezbędne w tworzeniu, a jedynie dodatkiem dla bardziej zaawansowanych użytkowników.

3.2.4 Przygotowanie klatek docelowych

W tym punkcie, analogicznie do wcześniejszych podobieństw pomiędzy filmem źródłowym i docelowym, wiele jest elementów wspólnych. Kroki w tworzeniu wideo docelowego są niemal takie same jak w przypadku zestawu danych źródłowych, jednak z kilkoma wyjątkami. Jednym z nich jest procesu ekstrakcji/wyrównywania twarzy (faces extraction/alignment proces).



Grafika 10 Identyfikacja dwóch twarzy na jednej klatce. Opracowanie własne.

Nadal dostępna jest metoda ekstrakcji ręcznej i S3FD, ale jest też taka, która łączy je obie i umożliwia zastosowanie specjalnego trybu ręcznego wyodrębniania. Ponadto folder „aligned_debug" jest generowany automatycznie (wynika to z możliwych błędów w dalszych fazach tworzenia). Dzięki temu, każdą usuniętą już w trakcie przeglądania twarz, można łatwo poprawić, używając funkcji „data_dst faceset manual re-extract deleted aligned_debug". Narzędzie pozwala poprawić obrysy twarzy, spośród tych usuniętych w trakcie przeglądania.

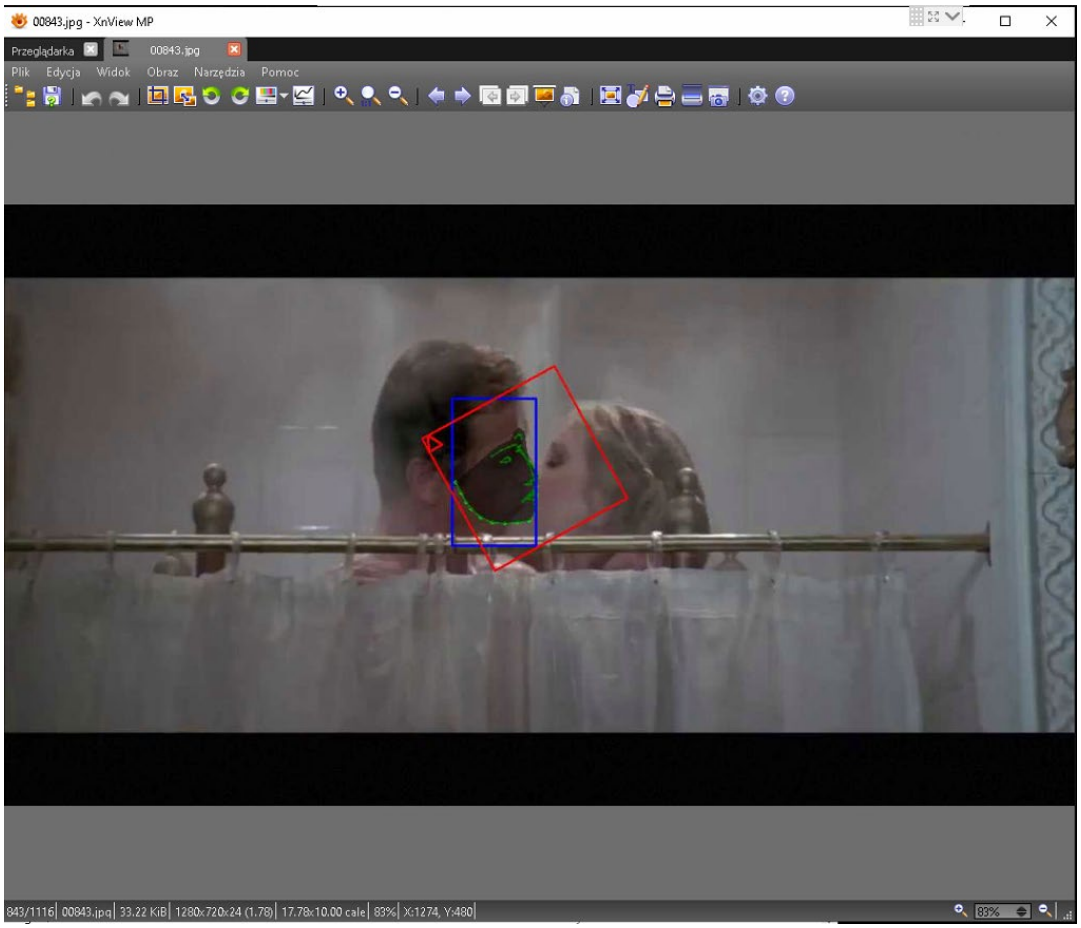
Ekstrakcja klatek

W poniższym procesie należy na podstawie przygotowanych wstępnie klatek wideo docelowego, wyodrębnić gotowe twarze. W narzędziu dostępne są takie ustawienia jak „data_dst faceset extract manual re-extract deleted aligned_debug", które pozwala na ręczne ponowne wyodrębnianie ramek usuniętych z folderu „aligned_debug". Jest również „data_dst faceset extract manual", który jest ekstraktorem ręcznym, podobnie jak w przypadku wideo źródłowych i pozwala na samodzielne wyodrębnianie twarzy. Narzędzie „data_dst faceset extract + manual fix" to

automatyczny, chociaż z możliwym ręcznym wprowadzaniem zmian, ekstraktor dla klatek, w których algorytm nie mógł poprawnie wykryć twarzy. Natomiast w pełni automatyczną ekstrakcję twarzy uruchamia narzędzie „data_dst faceset extract", które przeprowadza automatyczną ekstrakcję za pomocą algorytmu S3FD.

Podobnie jak w przypadku wideo źródłowego, dla „data_dst faceset extract" dostępne są 3 tryby ekstrakcji. Pełna twarz (WF) – dostępna do modeli trenowanych dla połowy (HF) i całej twarzy (FF). Cała twarz (FF) dostępna zarówno dla całej twarzy (WF) ale również współpracuje z innymi – HF, FF oraz z HEAD. Głowa (HEAD) służy do zamiany całej głowy i współpracuje wyłącznie z modelami HEAD. Nie nadaje się dla osób z długimi włosami oraz działa najlepiej, jeśli źródłowy zestaw twarzy pochodzi z jednego źródła, a zarówno nagranie źródłowe, jak i docelowe, mają krótkie włosy lub takie, które nie zmieniają kształtu w zależności od ruchu.

Wybór obszaru ekstrakcji zależy od typu twarzy modelu, który chce się trenować oraz wybranego formatowania twarzy w wideo źródłowym. Ważne jest również jaki rodzaj treningu będzie zastosowany w dalszej części. Dla przykładu model Quick 96 współpracuje wyłącznie z pełnymi twarzami (FF).



Grafika 11 Niedokładna identyfikacja jednej twarzy. Opracowanie własne.

Wyrównanie twarzy (aligned debug)

Folder „aligned debug" jest generowany automatycznie. Dostępne opcje są identyczne jak w przypadku wideo źródłowego. Pozwoli on dokładnie dostosować (*aligns*) dane źródłowe, takie jak twarz osoby, którą chce podmienić, do obrazu docelowego. W fazie wyrównywania twarzy (*aligned debug*) twórca dokładnie dostosowuje materiały źródłowe, aby ich pozycja, rozmiar i orientacja były zgodne z obrazem docelowym.

Czyszczenie klatek wideo docelowego

Po wyrównaniu twarzy (aligned debug data_dst), podobnie jak przy wideo źródłowym, należy je wyczyścić. Dostępny wybór metod sortowania, jest identyczny jak dla twarzy wyodrębnionych z nagrania źródłowego.

Samo czyszczenie docelowego zbioru danych nieco różni się jednak od źródłowego. Wynika to z tego, iż należy wszystkie twarze wyrównać dla wszystkich ramek, w których są obecne – w tym te zasłonięte. Pomocne może być przy tym kilka narzędzi, na przykład takich jak „data_dst view aligned results". Pozwala ono wyświetlić

zawartość folderu „aligned” za pomocą zewnętrznej aplikacji (wbudowanej w DFL), która oferuje szybsze odtwarzanie miniatur niż domyślny eksplorator zainstalowany w systemie Windows.

Narzędzie „data_dst view aligned_debug results” umożliwia szybkie przeglądanie zawartości folderu „aligned_debug” w celu zlokalizowania i usunięcia wszelkich ramek, w których twarz osoby docelowej ma niepoprawnie wyrównane punkty orientacyjne lub w których punkty orientacyjne nie zostały w ogóle umieszczone (co oznacza, że twarz w ogóle nie została wykryta). Używa się tego, aby sprawdzić, czy wszystkie twarze są prawidłowo wyodrębnione i wyrównane (jeśli punkty orientacyjne na niektórych ramkach nie pokrywają się z kształtem twarzy lub oczu / nosa / ust / brwi lub ich brakuje – należy je usunąć. W takim wypadku pod koniec procesu należy je ponownie ręcznie wyodrębnić lub wyrównać.

„Data_dst sort” podobnie jak w przypadku źródłowego zestawu danych, jest narzędziem umożliwiającym sortowanie wszystkich wyrównanych twarzy w folderze „data_dst/aligned”, dzięki czemu łatwiej jest zlokalizować nieprawidłowo wyrównane twarze, fałszywe alarmy i twarze innych osób, których nie należy trenować. „Data_dst util faceset pack” – podobnie jak w przypadku source – pozwala szybko spakować cały zestaw danych do jednego pliku. Dzięki temu można łatwiej przenosić lub udostępniać przygotowany zestaw twarzy. Narzędzie „data_dst util faceset unpack” rozpakowuje zestaw twarzy data_dst z jednego pliku przygotowanego poprzednim narzędziem. „Data_src util faceset resize” – pozwala zmienić jakość przechowywanego zbioru twarzy źródłowych. „Data_dst util recover original filename” podobnie jak w przypadku danych źródłowych, przywraca oryginalne nazwy/kolejność wszystkich wyrównanych klatek po ich przesortowaniu.

3.2.5 Trening modelu XSeg i znakowanie zestawu twarzy

Celem lepszego dopasowania twarzy jednej do drugiej, zalecane jest utworzenie oraz wytrenowanie modeli obu twarzy. Od listopada 2021 roku DFL udostępnił przeszkolone modele, które przy mniej skomplikowanych nagraniach (bez przedmiotów zasłaniających twarz takich jak okulary czy machanie rękoma) można łatwo stosować. Przetrenowany model uwzględnia charakterystyczne dla każdej twarzy elementy i pozwala przyspieszyć proces treningu, zwłaszcza jeżeli dotyczy tej samej lub podobnie wyglądającej twarzy.

Wbudowane narzędzie „XSeg Generic" udostępnia wstępnie przeszkolony ogólny model całych twarzy XSeg (lokalizacja: internal/model_generic_xseg). Funkcję tą zastosowano w przypadku tworzenia nagrania numer 2 (modelu żeńskiego influencerki). „XSeg Generic data_dst whole_face mask" stosuje maski całej twarzy do zestawu danych docelowych, natomiast „XSeg Generic data_src whole_face mask" pozwala zastosować maski całej twarzy do zestawu danych źródłowych.

Niektóre typy twarzy wymagają jednak zastosowania innej maski niż domyślna, dostępna poprzez narzędzie Xseg Generic. Internauci dzielą się nimi sowich repozytoriach, dzięki czemu bardziej doświadczona osoba może zastosować inny model.

Najczęściej dostępne modele implikuje się od razu z zestawem danych po jego wyodrębnieniu – domyślne maski pochodzą z punktów orientacyjnych i obszaru pokrycia podobnego do typu całej twarzy (WF). Konieczne jest stosowanie gotowych modeli Xseg w przypadku trenowania głowy. Natomiast gotowych modeli nie stosuje się w przypadku pełnej lub połowy twarzy. Maski XSeg są również wymagane w przypadku, gdy planujemy użyć funkcji Face lub Background Style Power (FSP, BGSP) podczas treningu modeli SAEHD/AMP. Stosujemy je wówczas niezależnie od wybranego typu twarzy nagrania źródłowego lub docelowego.

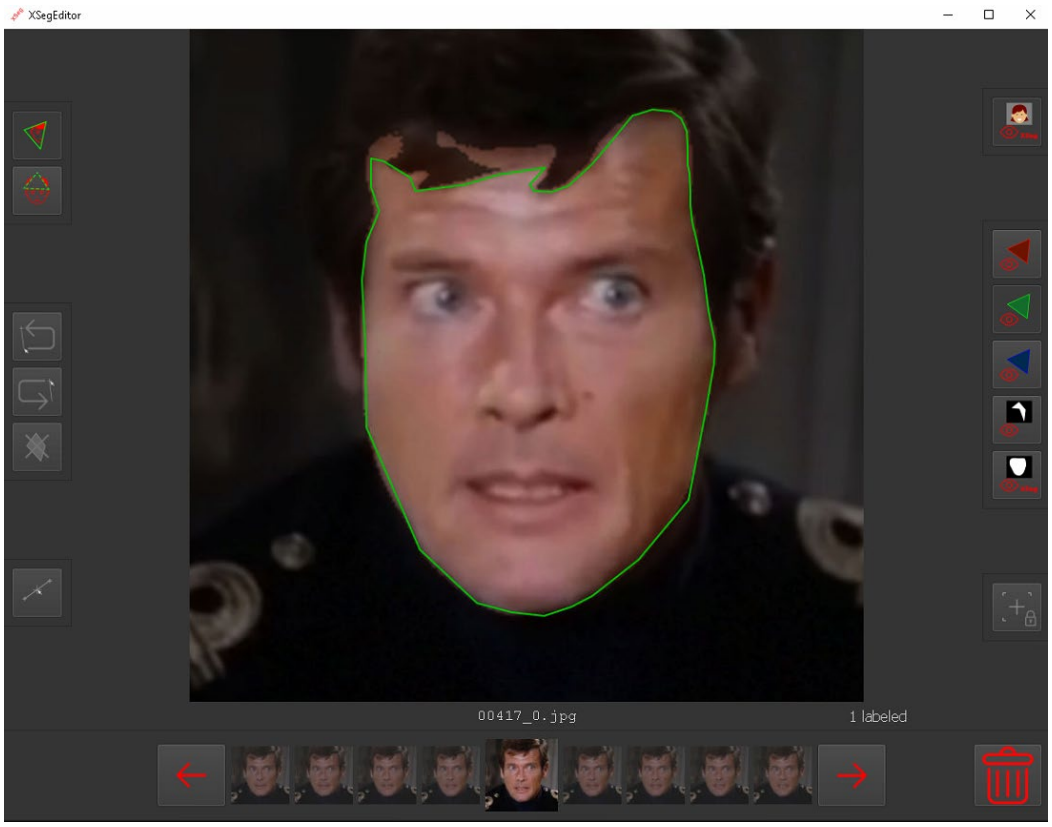
W dużym uproszczeniu XSeg pozwala określić, w jaki sposób chce się maskować twarze oraz które części twarzy będą trenowane, a które nie. Pozwala to na wykluczenie z treningu przeszkód w obszarze twarzy, takie jak ręka, mikrofon, lizak itd. Przedmioty te dzięki utworzeniu maski można w treningu wykluczyć, na skutek czego podczas trenowania SAEHD elementy te nie będą zakłócać trenowanego obrazu.

XSeg pozwala wykluczyć prawie wszystkie przeszkody, takie jak dłonie, palce, kolczyki, blizny, tatuaże na twarzy, pojedyncze kosmyki włosów, a nawet okulary. Poniżej przedstawiono najważniejsze funkcje, które mogą być przydatne przy dalszej pracy. „XSeg model", to model wyszkolony przez użytkownika, używany do nakładania masek na zestawy danych SRC i DST, a także do nakładania twarzy podczas procesu scalania. „XSeg label" tworzy pole, które rysuje się na twarzy, aby zdefiniować jej obszar i dostarczyć modelowi informacji które elementy powinien trenować, a które nie. „XSeg mask" to maska wygenerowana i zastosowana do zestawu danych SRC lub DST przez dostarczony wcześniej i przeszkolony model XSeg. Natomiast „XSeg dataset" to zbiór oznaczonych twarzy (tylko jeden określony typ lub zestaw danych SRC lub DST, oznaczonych w podobny sposób). Są one często udostępniane na forum przez

użytkowników i są świetnym sposobem na rozpoczęcie tworzenia własnego zestawu, ponieważ można go pobrać i całego wykorzystać, lub wybrać z niego konkretne twarze, których potrzeba (albo dodać do niego własne obrobione twarze, które są oznaczone w wyodrębnione w podobny sposób).

Maski określają, który obszar na próbce twarzy jest samą twarzą, a co tłem lub przeszkodą. W przypadku danych źródłowych oznacza to, że cokolwiek zostanie włączone do maski, zostanie przeszkolone przez model z wyższym priorytetem, podczas gdy wszystko inne zostanie przeszkolone z niższym priorytetem (lub mniejszą precyzją). W przypadku danych docelowych, jest tak samo – maska pozwala wykluczyć przeszkody, aby model nie traktował ich jako części twarzy, a także aby później podczas łączenia, te przeszkody (na przykład lizak) były widoczne i nie zostały zakryte przez ostatecznie utworzoną twarz.

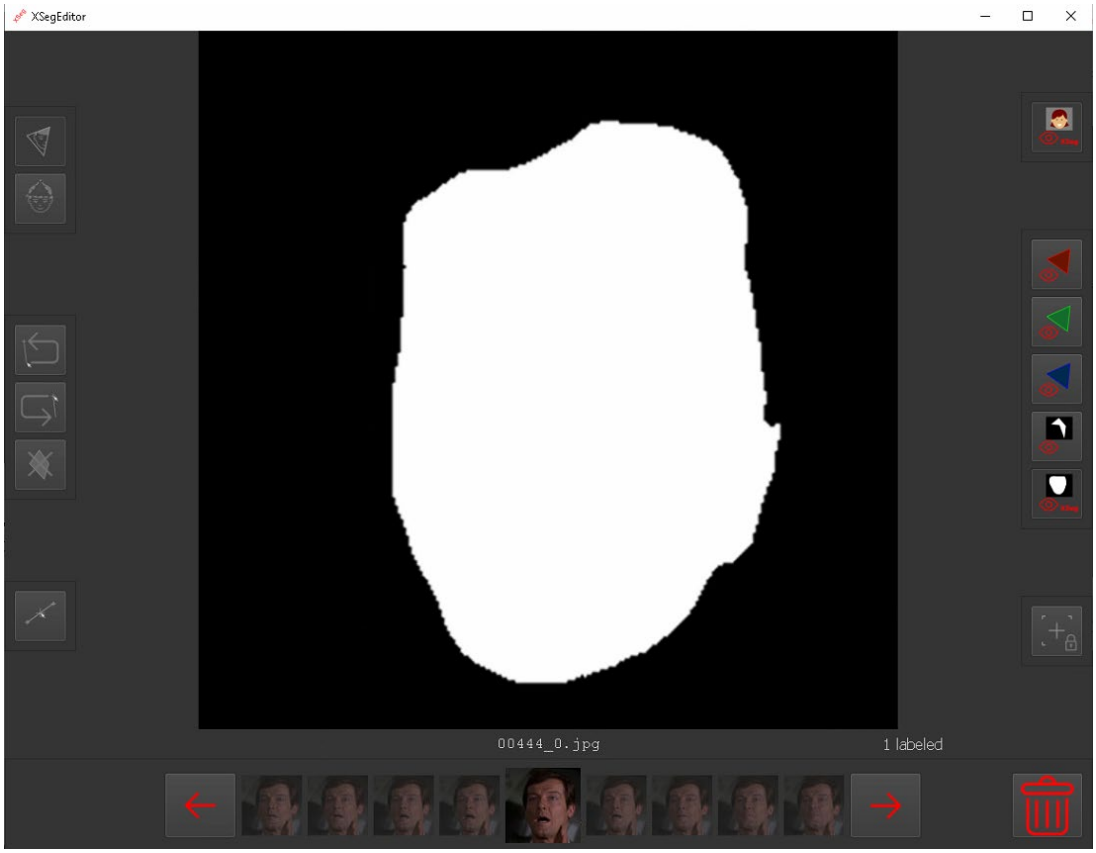
Przy używaniu XSeg dostępne są wspierające modelowanie funkcje. „Data_dst mask for XSeg trainer – edit" to narzędzie do oznaczania granic obszaru twarzy w nagraniu docelowym, za pomocą wielokątów XSeg. Pozwala ono samodzielnie tworzyć XSeg label dla twarzy docelowych (tworzy obszary robocze twarzy – grafika 12). „Data_dst mask for XSeg trainer – fetch" pozwala utworzyć kopię oznaczonych twarzy DST do folderu „aligned_xseg" celem ich archiwizacji lub dalszej obróbki. Narzędzie data_dst mask for XSeg trainer – remove" usuwa etykiety z twarzy docelowej.



Grafika 12 Weryfikacja zarysowań twarzy. Opracowanie własne.

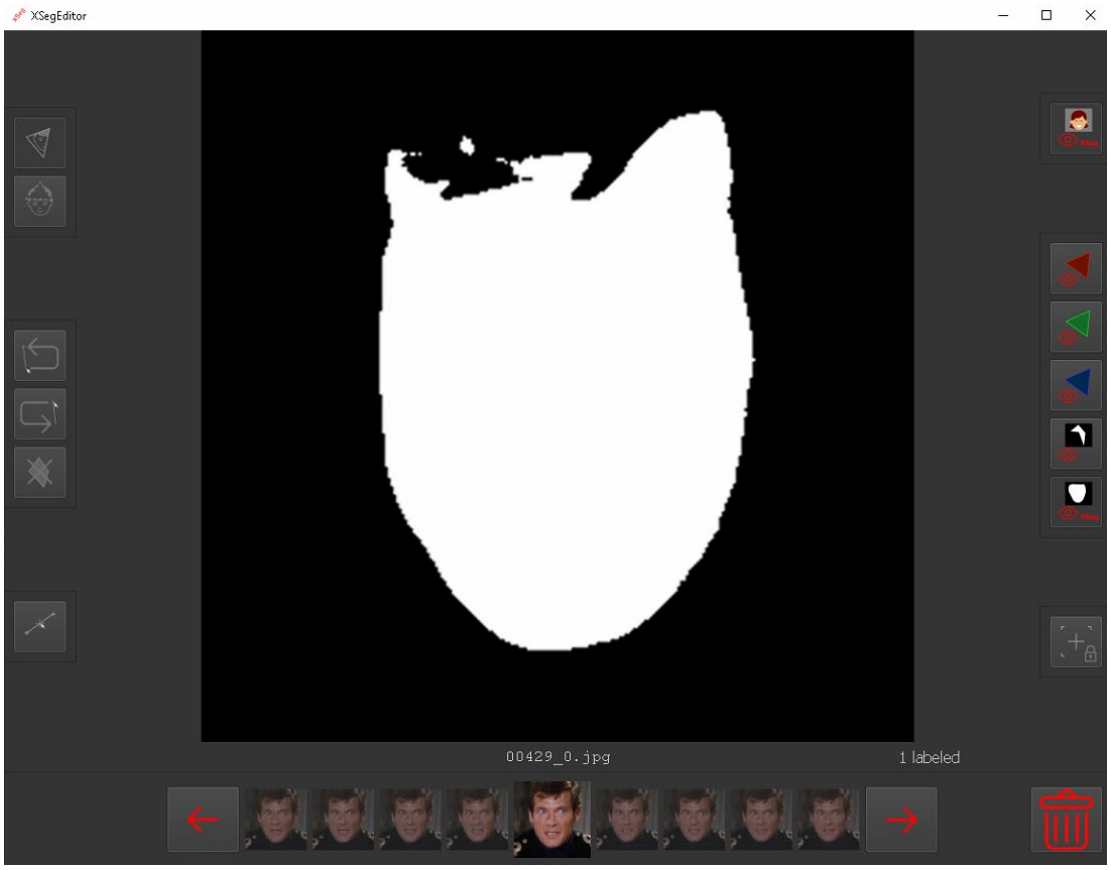
Analogiczne funkcje dostępne są dla nagrania źródłowego. „Data_src mask for XSeg trainer – edit" to narzędzie do oznaczania granic obszaru roboczego wideo źródłowego wielokątami XSeg. Pozwala samodzielnie tworzyć XSeg label dla wideo źródłowego (obszary robocze twarzy – grafika 12). „Data_src mask for XSeg trainer – fetch" tworzy kopię oznaczonych twarzy źródłowych do folderu „aligned_xseg", natomiast „data_src mask for XSeg trainer – remove" usuwa etykiety z twarzy źródłowych.

Interfejs narzędzia „XSeg data_src mask" jest niezwykle rozbudowany. Z lewej strony przyciski pozwalają na dokonywanie zmian w opracowanym obszarze roboczym twarzy (XSeg label). Przyciski umieszczone z prawej strony dają możliwość tworzenia obszaru roboczego twarzy jak i dodawanie / odejmowanie elementów które chce się zawrzeć lub wykluczyć.



Grafika 13 Sprawdzanie wyciętej maski. Opracowanie własne.

Powyżej (grafika 13) prezentowany jest roboczy obszar, wytrenowany automatycznie przez XSeg twarzy aktora. Jak można zauważyć wykluczone zostały z niego obszary zacieniowane i te przykryte włosami (tam, gdzie twarz nie została wykryta). W przypadku grafiki 14. wyznaczonych zostało o wiele więcej miejsc zasłanianych przez włosy. Celem osiągnięcia dobrego efektu końcowego, zalecana jest ręczna edycja takiej maski. Chęć uwzględnienia ominiętych obszarów wymaga indywidualnej korekty kilku różniących się od siebie obrazów oraz ponowne uruchomienie skanowania masek.

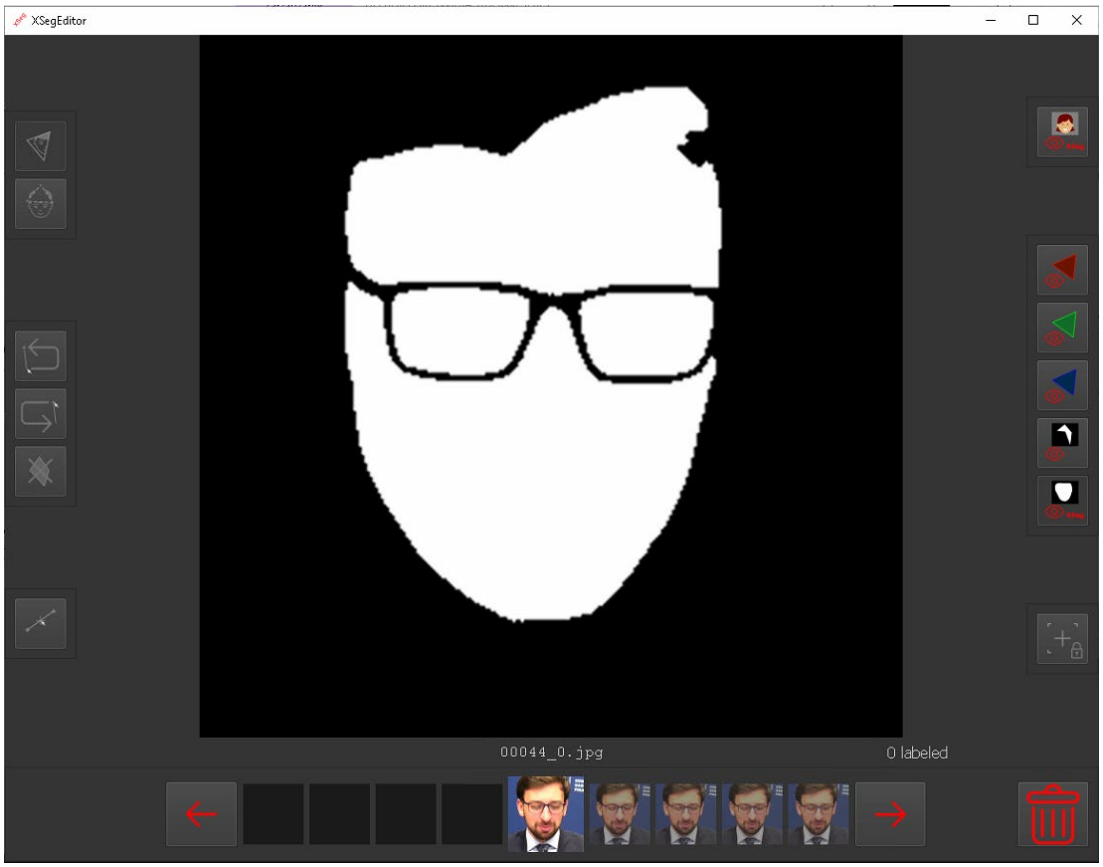


Grafika 14 Poprawianie wyciętej maski. Opracowanie własne.

Ostatnia funkcja – „XSeg) train.bat” – rozpoczyna trenowanie modelu XSeg. Trenowanie odbywa się jednocześnie dla modeli DST jak i SRC. Czym więcej obrazów zostanie załadowanych ręcznie do modelu oraz czym dokładniejsze one będą tym lepiej dla późniejszego dopasowywania twarzy.

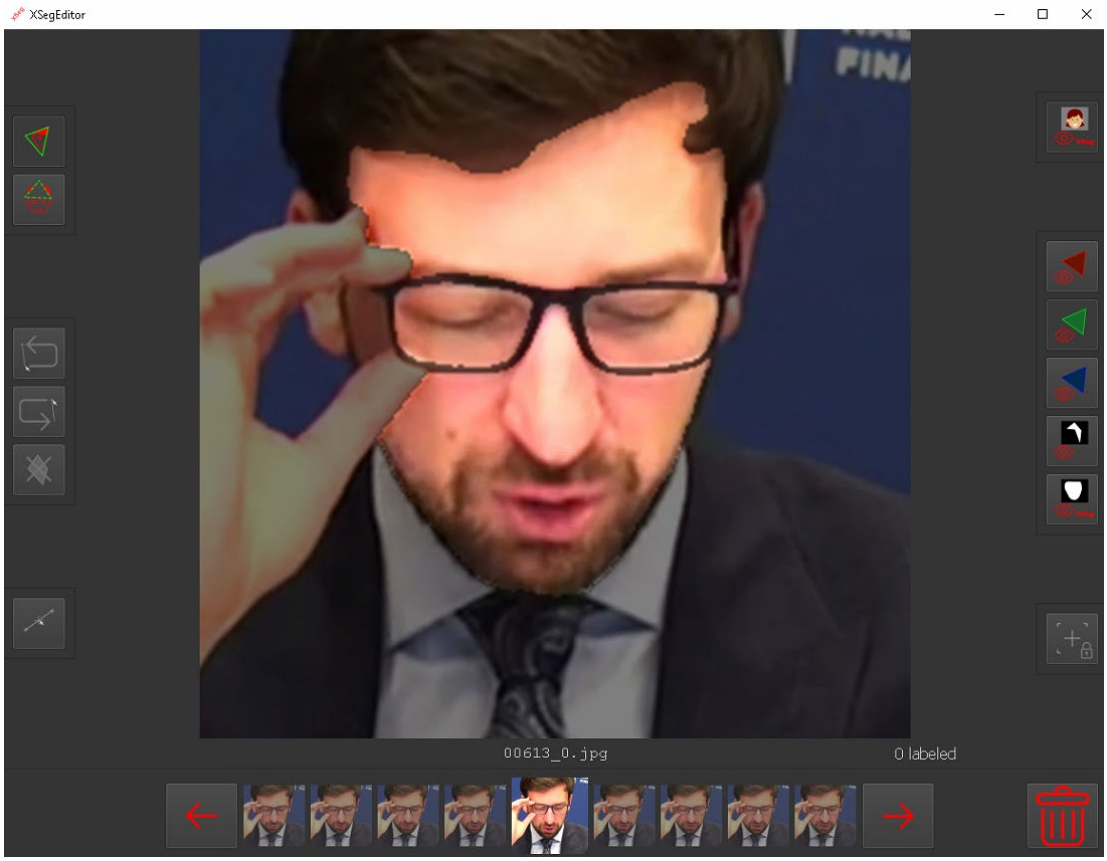
Grafika 15 Usuwanie elementów zasłaniających maskę. Opracowanie własne.

W trakcie trenowania maski XSeg dostarcza szeregu funkcji mogących usprawnić proces. „Trained mask for data_dst – apply" generuje i stosuje maski gotowe XSeg do twarzy docelowych. Funkcja „trained mask for data_dst – remove" usuwa zastosowane maski XSeg i przywraca domyślne maski całej twarzy. „Trained mask for data_src – apply" generuje i stosuje maski XSeg do twarzy docelowych, natomiast „trained mask for data_src – remove" usuwa maski XSeg i przywraca domyślne ustawienia maski.



Grafika 16 Okulary jako element zasłaniający twarz. Opracowanie własne.

Dla uzyskania lepszego efektu sugeruje się kilkukrotnie naprzemiennie wykonanie wpierw ręcznego oznaczenia kilkudziesięciu obszarów roboczych twarzy dla danych docelowych oraz danych źródłowych. Czym bardziej odmienne będą od siebie ręcznie oznaczane twarze, tym większa szansa na lepsze trenowanie modelu. Następnie przetrenowanie modelu do ok. 20k iteracji. Następnie zaleca się zaimplikowanie modelu dla danych docelowych i danych źródłowych, by następnie ponownie dokonać ręcznych poprawek (wykluczyć obszary źle rozpoznane lub dodać te które model wykluczył). Sugeruje się kilkukrotnie powtórzyć tę czynność, tak by utworzone maski miały jak największą dokładność.



Grafika 17 Ręczne wycinanie ręki z maski. Opracowanie własne.

3.2.6 Trening deepfake

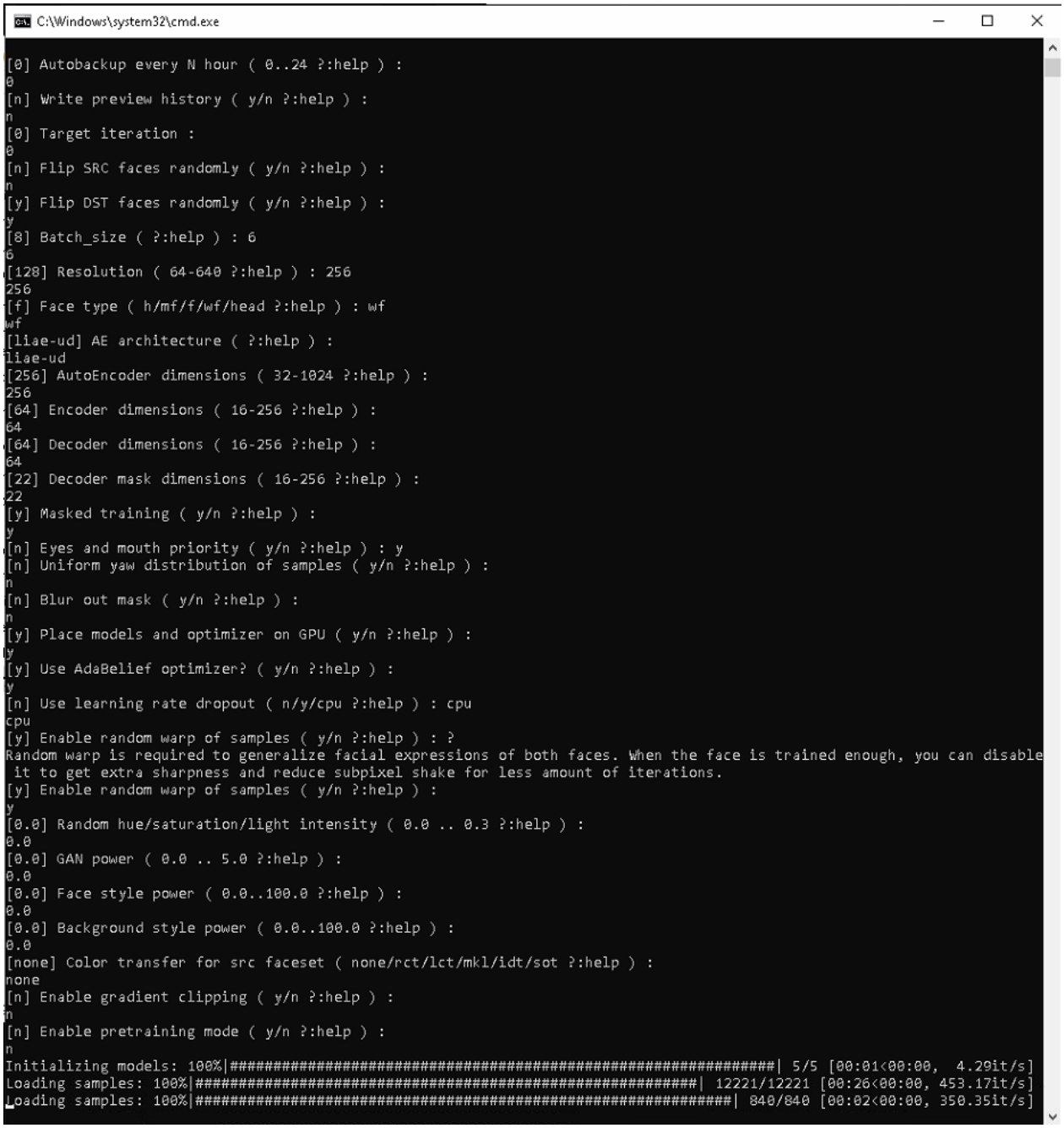
Do trenowania modeli dostępne są 3 narzędzia: SAEHD oraz mniej zaawansowany i prostszy w swojej obsłudze Quick96 i AMP – ciągle w fazie testów i doskonalenia.

Model SAEHD wymaga minimum 6 GB VRAM. Pozwala na swobodną modyfikację parametrów i oferuje zaawansowane opcje. Wybrany został jako model treningowy dla niniejszego badania.

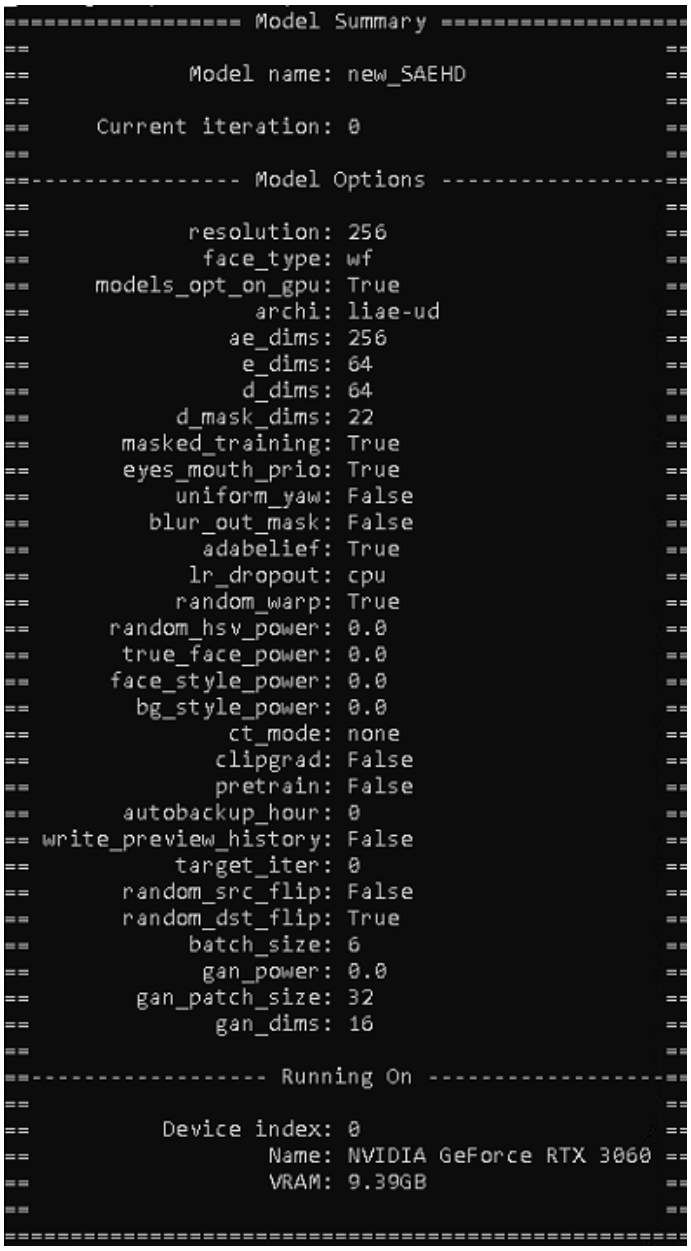
Model Quick96 obsługuje karty graficzne z pamięcią już od 2 GB. Jest to prosty tryb, często nazywany testowym, dedykowany dla słabszych kart graficznych. Posiada stałe parametry ustawień, których nie można modyfikować. Rozdzielczość trenowanych twarzy to 96x96 pixeli, jedyny rodzaj twarzy masek to cała twarz, batch size wynosi 4, a szkolenie odbywa się w architekturze DF-UD.

Model trenowania AMP jest najnowszym z dostępnych. Obsługuje karty z minimalną pamięcią 6 GB VRAM. Jest to nowy typ modelu, który wykorzystuje inną architekturę, zmienia kształty (próbuje zachować kształty twarzy źródłowej, delikatnie

modelując twarz docelową). Ponadto, w przeciwieństwie do treningu SAEHD, posiada możliwość regulowania współczynnika morfingu (uczenia i łączenia).

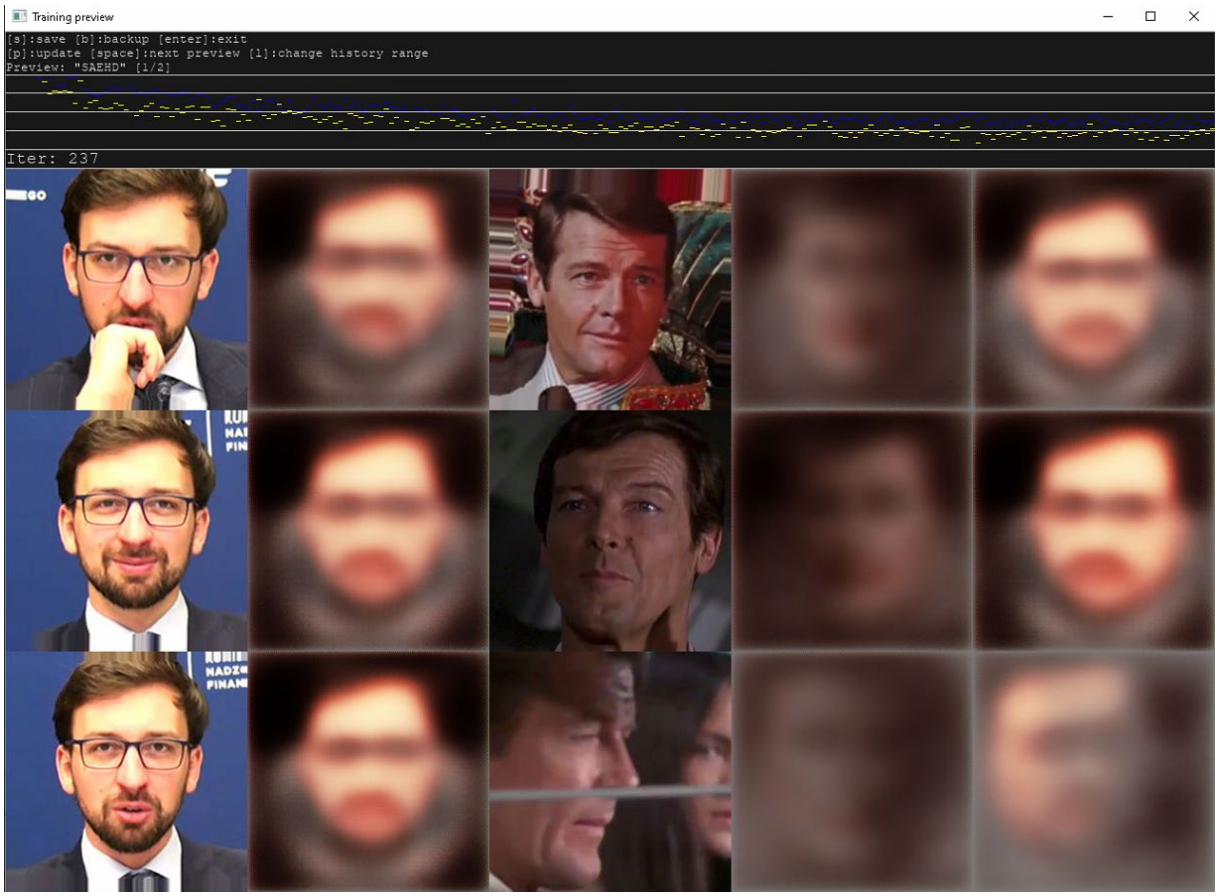


Grafika 18 Konfigurowanie treningu SAEHD. Opracowanie własne.



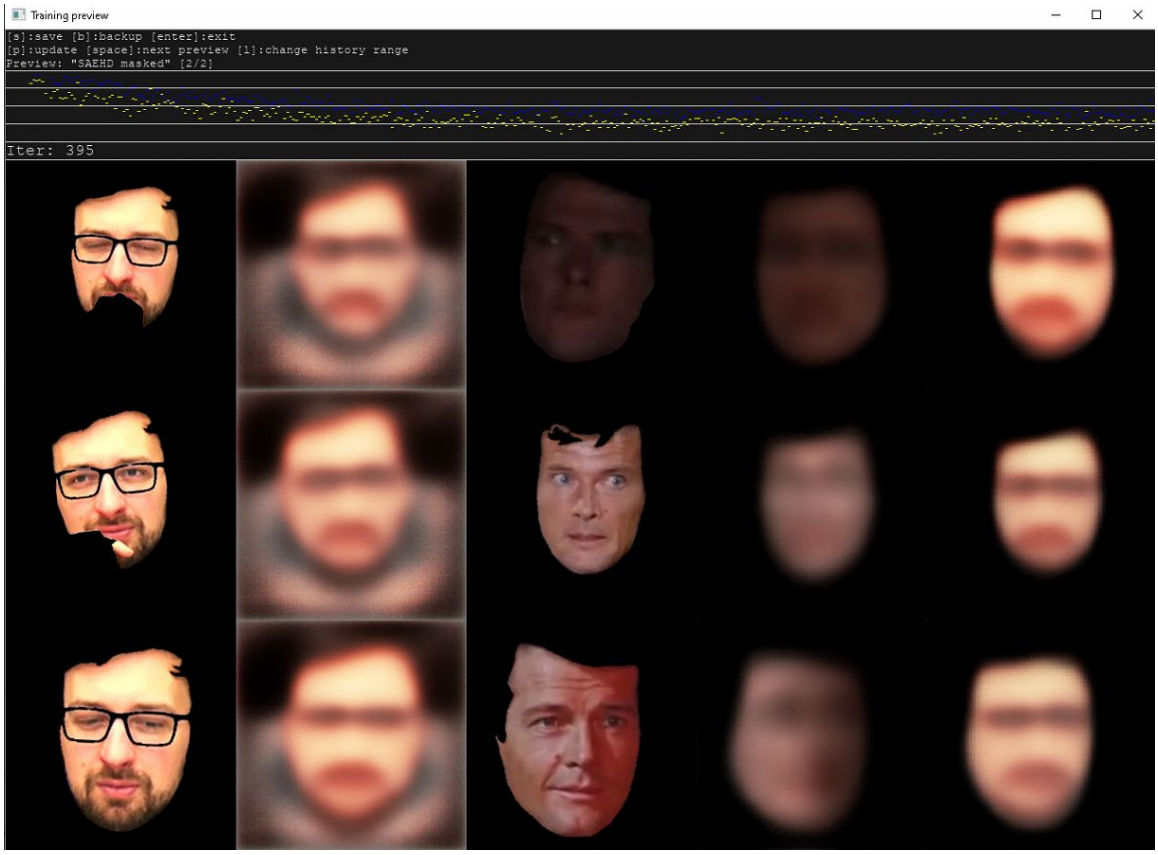
Grafika 19 Zastosowane ustawienia treningu SAEHD. Opracowanie własne.

Trenowanie modelu odbywa się na podstawie działania przeciwstawnych sieci neuronowych. Skrypt tworzy model twarzy na podstawie dostępnych danych, a następnie weryfikuje otrzymany obraz z oryginalnym. Po zweryfikowaniu niezgodnych elementów następuje próba go polepszenia, poprzez kolejną iterację treningową. Czynność ta przebiegać może w nieskończoność, jednak jej efekty wraz z kolejnymi iteracjami będą coraz mniejsze, a różnica pomiędzy efektem początkowym, a końcowym coraz mniejsza.



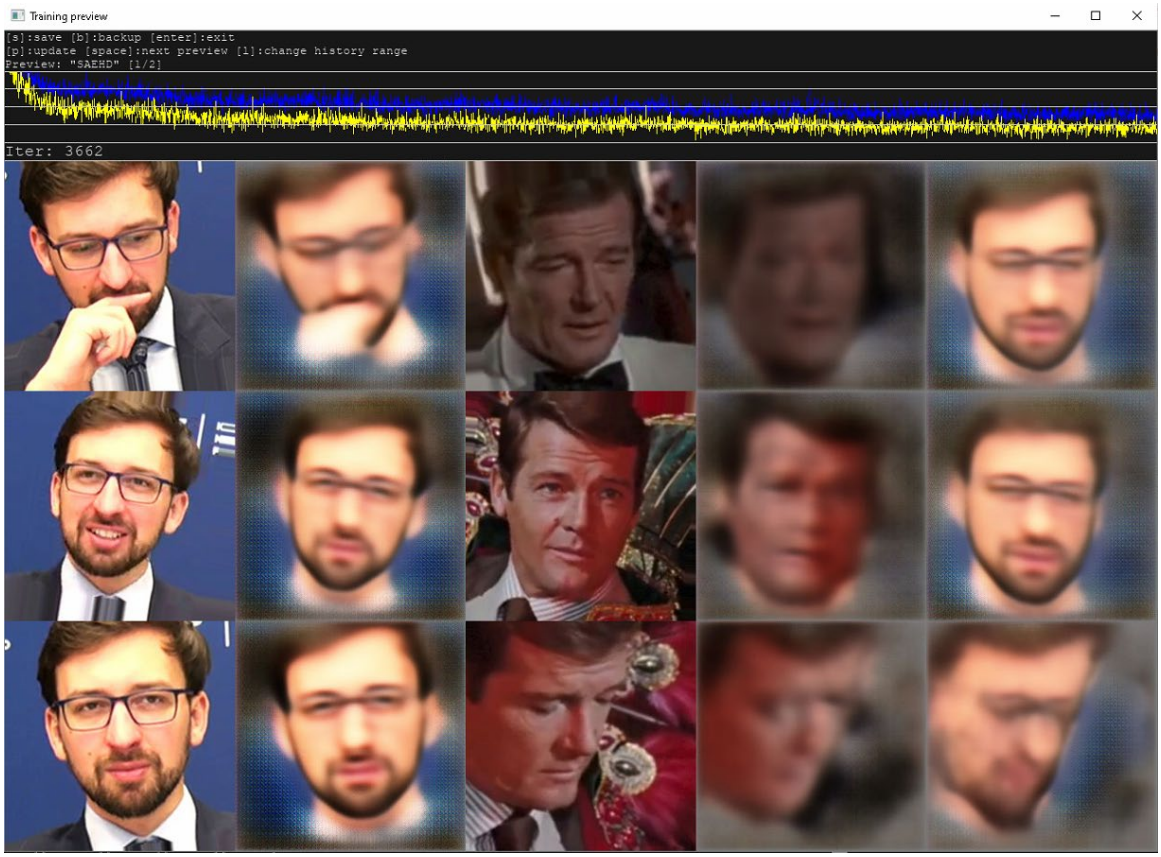
Grafika 20 Początkowe postępy treningu – widok całej twarzy. Opracowanie własne.

W pierwszej kolumnie widoczne są twarze wyodrębnione z wideo źródłowego, zaś w drugiej generuje się na tej podstawie twarz. Podobnie dzieje się z kolumną trzecią i czwartą, gdzie wpierw widoczne są kadry twarzy z filmu, a następnie trenowana na ich podstawie nowa maska. W ostatniej kolumnie powstaje docelowa maska – obraz źródłowy po zweryfikowaniu obrazu docelowego.



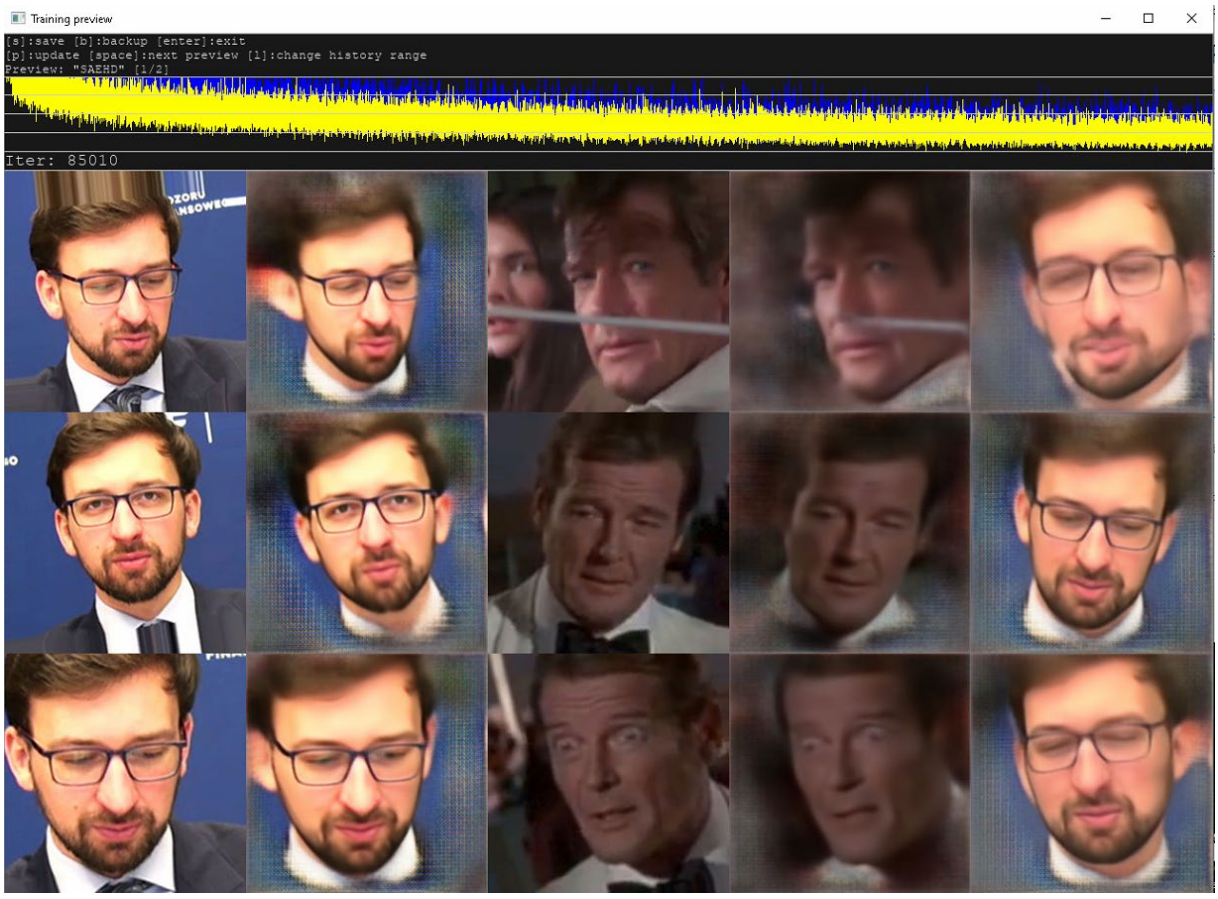
Grafika 21 Początkowe postępy treningu – widok maski. Opracowanie własne.

Powyższy screen obrazuje wyłącznie maski twarzy, które obecnie podlegają treningowi. Na powyższym przykładzie wyraźnie widać źle opracowaną w xSeg maskę źródłową (drugi kadr, pierwsza kolumna). Widoczny palec jest w tym momencie rozpoznawany przez SAEHD jako część twarzy i może zaburzyć efekt końcowy. Dla uzyskania lepszej jakości, po zidentyfikowaniu błędu, należy przerwać pracę i poprawić zidentyfikowaną maskę.



Grafika 22 średni etap treningu – widok całej twarzy. Opracowanie własne.

Już w trakcie treningu zauważyć można jak dużą trudność sprawiają modelowi dodatkowe elementy zasłaniające twarz, takie jak np. okulary (nawet w przypadku ich prawidłowego usunięcia przy tworzeniu maski). Pomimo ich wykluczenia z treningu, w dalszym stopniu może się zdarzyć, że ich fragmenty będą miały wpływ na efekt końcowy.



Grafika 23 Zaawansowany etap treningu – widok całej twarzy. Opracowanie własne.

Jak można zaobserwować na powyższym przykładzie, pomimo 85k iteracji i włączonego priorytetu dla ust i oczu, modelowi SAEHD nadal ciężko jest umiejscowić spojrzenie w odpowiednim kierunku.

3.2.7 Scalanie obrazów

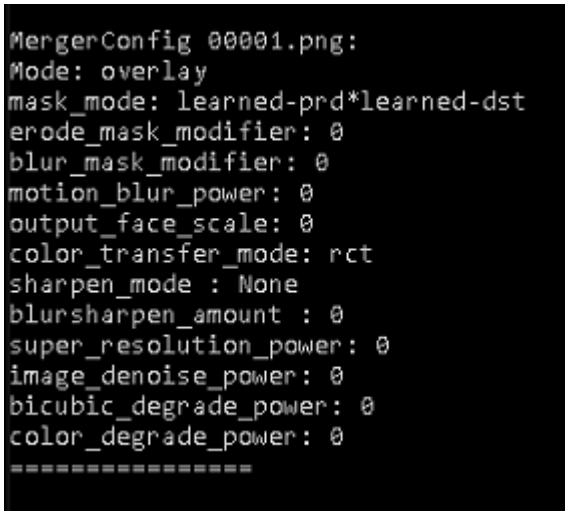
Po zakończeniu trenowania należy nanieść wyuczoną twarz na oryginalne klatki filmowe. Do tej czynności dostępne są 3 konwertery odpowiadające 3 dostępnym modelom:

- Merging SAEHD,
- Merging AMP,
- Merging Quick96.

Najlepiej wybrać ten sam konwerter, który był wykorzystywany do treningu. Po wybraniu dowolnego z nich pojawi się okno wiersza poleceń z kilkoma komunikatami.

Pierwszy pyta czy chce się użyć interaktywnego konwertera, więc domyślną wartością jest y (włączony) i zaleca się używanie go zamiast zwykłego, ponieważ ma

wszystkie funkcje, a także interaktywny podgląd, w którym widać efekty wszystkich zmian nanoszonych przy edycji opcji i włączaniu/wyłączaniu różnych funkcji.



Grafika 24 Ustawienia końcowe scalania. Opracowanie własne.

Powyżej (grafika 24) zaprezentowane są ustawienia, które poprzez interaktywny podgląd można swobodnie zmieniać. Służą one przede wszystkim do wygładzania konturów maski i lepszego dopasowywania scalanych elementów. W zależności od wybranego treningu, poszczególne ustawienia pozwalają na uzyskanie bardziej zadowalającego efektu.

Aplikacja oferuje kilka trybów (mask_mode) nakładania nowej twarzy na klatki filmowe. Najważniejszym jest funkcja nakładania (overlay), która dokonuje prostego nałożenia wyuczonej twarzy na klatkę. Funkcja „original" wyświetla oryginalną klatkę bez zamienionej twarzy. „Hist-match" nakłada wyuczoną twarz i dostosowuje ją w oparciu o histogram. Funkcja ta posiada 2 tryby: normalny i maskowany, które można przełączać, jednak zalecane jest wybranie normalnego trybu. Funkcja „seamless" wykorzystuje funkcję „opencv poisson", aby połączyć nową wyuczoną twarz nad głową w oryginalnej ramce. Tryb ten może powodować migotanie obrazu i niedopasowanie twarzy. „Seamless hist match" łączy w sobie zarówno dopasowanie histogramowe, jak i "seamless", natomiast tryb „raw-rgb" nakłada surową wyuczoną twarz bez maskowania.

Wybranie trybu nakładania ma wpływ na pozostałe funkcje dostępne w aplikacji, w trakcie nakładania. „Hist match threshold": kontroluje siłę dopasowania histogramu w trybie dopasowania histogramowego oraz „seamless hist match". Wartość można zmniejszyć lub zwiększyć.

Tryb „erode mask” pozwala łatwo kontrolować rozmiar maski, którą można dowolnie pomniejszać lub powiększać, tak by dopasować ją do oryginalnej twarzy. Funkcja rozmywania („blur”) maski pozwala rozmyć lub wyostrzyć krawędzie maski, celem płynniejszego przejścia twarzy. Efekt rozmywania można zmniejszyć lub zwiększyć w zależności od oczekiwanych efektów.

Funkcja rozmywania ruchu („motion blur”) jest bardziej zaawansowaną funkcją od zwykłego „blur”. Po wprowadzeniu parametrów początkowych funkcja ładuje wszystkie klatki i oblicza wektory ruchu twarzy. Następnie tworzy efekty rozmycia ruchu, które dodaje w miejscach, w których twarz się porusza. Wysoka wartość tego ustawienia może spowodować, iż nawet niewielki ruch może ulec rozmyciu.

„Super resolution” umożliwia wyostrzenie takich elementów twarzy jak zęby, oczy i usta. Ich tekstura staje się wówczas wyraźniejsza i pozwala wyeksponować więcej ich szczegółów. Podobnie działa funkcja „Blur/sharpen”, która rozmywa lub wyostrza wyuczoną twarz. Działa w dwóch trybach – „box” lub „gaussian”.

Skalowanie twarzy pozwala na powiększenie lub pomniejszenie wytrenowanej twarzy, zaś „mask modes” modyfikuje tryby maskowania. W zależności od potrzeb może dopasować wyląd maski do twarzy docelowej lub tylko do maski docelowej, wyuczonej podczas treningu. Może również dostosować się do maski źródłowej lub obu masek jednocześnie. Pozostałe tryby pozwalają zwiększyć lub zmniejszyć wpływ jednej lub drugiej maski lub twarzy. Tryb „learned-prd*dst*XSeg-dst*prd” łączy w sobie wszystkie 4 tryby, gdzie każdy z nich uwzględniany jest przy dopasowywaniu twarzy.

Funkcja „color transfer modes” podobnie jak w przypadku transferu kolorów podczas treningu, pozwala lepiej dopasować kolor skóry wyuczonej twarzy do oryginalnej twarzy. Służy to uzyskaniu bardziej płynnej i realistycznej zamianie twarzy. Dostępnych jest 8 różnych trybów, które kopiują kolory przy zastosowaniu odmiennych filtrów

Funkcja „image degrade modes” dotyczy ustawień całej klatki. Zmieniając ją, można dowolnie odszumić obraz czyniąc go lekko rozmytym, zmniejszyć głębię koloru lub rozmyć.

Każde z tych ustawień może pomóc w zamaskowaniu niedoskonałości treningu i ulepszyć efekt końcowy. Profesjonaliści końcowym etapem określają mozolne, poklatkowe przeglądanie nagrania i ręczne dopracowywanie szczegółów w aplikacjach graficznych, takich jak Adobe After Effects.

3.3 Wnioski – powstawanie nagrań deepfake obecnie oraz w przyszłości

Niniejszy rozdział szczegółowo opisuje najpopularniejszą obecnie metodę tworzenia nagrań deepfake. Zdaniem autorów aplikacja DeepFaceLab jest wykorzystywana przy tworzeniu ponad 95% filmów[239]. Autorzy oprogramowania szczycą się, że przy pomocy ich aplikacji powstawały filmy takich twórców jak deeptomcruise, 1facerussia, arnoldschwarzneggar, mariahcareyathome?, diepnep, mr__heisenberg, deepcaprio, VFXChris, Ume, Sham00k, Collider videos, iFake, NextFace, Futuring Machine, RepresentUS, Corridor, Crew, DeepFaker, DeepFakes in movie, DeepFakeCreator, Jarkan i wielu innych[240]. Nagrania dostępne na tych kanałach mają obecnie miliardy wyświetleń oraz miliony polubień. Tylko jeden kanał 1facerussia, parodiujący w swoich nagraniach prezydenta Federacji Rosyjskiej – Władimira Putina, na początku 2023 roku zgromadził ponad 100 milionów polubień i ponad 11 milionów obserwujących[241].

Jak można zauważyć, tematyka powyższych kanałów dotyczy głównie szeroko pojętej humorystyki. Jest to obecnie drugi (zaraz po pornografii) najpopularniejszy temat przeróbek deepfake. Coraz częściej nagrania deepfake powstają jednak w odmiennych celach, takich jak chęć oszustwa czy manipulacji. W większości przypadków nagrania mają na celu zdyskredytowanie znanych osób lub obniżenie zaufania do nich. Wielokrotnie próbowano również przy pomocy nagrań deepfake wpłynąć na świadomość społeczną, by bezpośrednio zagrozić bezpieczeństwu państwa. Swój wpływ nagrania deepfake mają również na bezpieczeństwo finansowe – oszuści coraz chętniej sięgają po tę technologię, próbując wykorzystać ją do kradzieży pieniędzy.

Pytanie badawcze postawione w niniejszym rozdziale brzmi: „w jaki sposób powstają nagrania deepfake?". Próbą odpowiedzi na nie była robocza hipoteza brzmiąca następująco: nagrania deepfake powstają poprzez wykorzystanie oprogramowania do zmiany obrazu lub dźwięku w oryginalnym nagraniu, tak, aby wyglądało to jakby ktoś inny mówił lub wyglądał inaczej niż w rzeczywistości.

---

[239] Strona główna oprogramowania DeepFaceLab, https://github.com/iperov/DeepFaceLab [dostęp: 01.01.2023].
[240] Tamże.
[241] Profil 1facerussia na TikTok-u, zawierający nagrania parodiujące Władymira Putina, https://www.tiktok.com/@1facerussia [dostęp: 01.01.2023].

W toku analizy zebranego materiału empirycznego oraz badań terenowych (samodzielne przygotowanie nagrań deepfake) hipotezę tę udało się zweryfikować pozytywnie. Profesjonalne tworzenie nagrania deepfake wymaga użycia technologii głębokiego uczenia maszynowego oraz wiedzy technicznej. Jak wykazano w niniejszym rozdziale, proces ten polega na użyciu gotowej aplikacji do przetwarzania oryginalnego nagrania, a następnie wykorzystania tego przetworzenia do stworzenia nowego nagrania, na którym twarz oryginalnego użytkownika zostaje zastąpiona twarzą innej osoby. Proces ten składa się z elementów takich jak przygotowanie danych, ekstrakcja obrazów, trenowanie maski, uczenie modelu sieci neuronowej, renderowanie oraz postprodukcja.

Wykazane zostało, iż na poziomie rozwoju technologii pod koniec 2021 roku (w tym okresie przygotowano materiały do badania) tworzenie nagrań deepfake było procesem skomplikowanym i wymagającym specjalistycznej wiedzy i umiejętności w zakresie obróbki graficznej i uczenia maszynowego, a także dostępu do specjalistycznego oprogramowania i komputerów wyposażonych w silne karty graficzne, tak aby móc przetwarzać duże ilości danych. Tworzenie nagrań deepfake nie było wówczas łatwo dostępne. Zwykły użytkownik, chcąc tworzyć takie filmy, musiał posiąść wpierw specjalistyczną wiedzę i zaopatrzyć się w niezbędne zasoby, a i tak wytwarzane przez niego nagrania często od razu weryfikowane były jako fałszywe.

Odmienna sytuacja nastąpiła w 2023 roku. Technologia deepfake staje się wówczas coraz popularniejsza, a podstawowe aplikacje dostępne są przez przyjazny interfejs graficzny. Upowszechniło się również tworzenie nagrań deepfake za opłatą, gdzie za niewielkie kwoty otrzymać można gotowy film. Współcześnie, gdy ktoś chce zorganizować profesjonalne nagranie, w Internecie odnaleźć można setki poradników, również w postaci wideo, ukazujących jak otrzymać zadowalający efekt.

DeepFaceLab, podobnie jak FaceSwap, jest narzędziem skierowanym do bardziej zaawansowanych użytkowników, którzy posiadają specjalistyczną wiedzę z zakresu uczenia maszynowego oraz pracy z grafiką. Wymaga ono pewnej ilości czasu i wiedzy, aby przejść przez wszystkie etapy i uzyskać satysfakcjonujący rezultat, jednak powoli aplikacje te wypierane są przez bardziej przyjazne aplikacje web-owe. Współcześnie jednak nawet aplikacje na telefon są w stanie wytworzyć kilkunastosekundowe nagranie deepfake, w jakości porównywalnej do nagrań przygotowywanych na potrzeby badania przez autora niniejszej pracy. Przykładową aplikacją może być aplikacja Face Swap

Video by Deep Fake, umożliwiającą utworzenie prostego nagrania deepfake o długości do 10 sekund, za darmo, w kilkanaście sekund, bezpośrednio z pozycji telefonu[242].

Przewiduje się, iż w przyszłości technologie sztucznej inteligencji i uczenia maszynowego będą dalej się rozwijać i stawać się coraz bardziej dostępne. Oznacza to, że tworzenie nagrań deepfake może być jeszcze łatwiejsze i bardziej powszechne. Współcześnie część badaczy uważa, iż kolejna ewolucja czatów sztucznej inteligencji (jak czat GPT-5) będzie w stanie samodzielnie tworzyć takie nagrania[243]. Propagatorzy nauki spekulują, iż powstające GPT-5 miałoby być nawet 500 razy większe od GPT-3.5 i zawierałby liczbę połączeń pomiędzy neuronami podobną jak w ludzkim mózgu – ponad 100T[244]. Zakładane jest, że w przyszłości tworzenie fałszywych nagrań uproszczone zostanie do wskazania materiałów źródłowych i oczekiwań końcowych, bez żmudnych treningów i odręcznych poprawek.

Zdaniem autora dysertacji, powszechny i coraz łatwiejszy dostęp do wyżej opisanych technologii stwarza warunki do skorzystania z pokusy wpływania na rzeczywistość społeczną w różnych sferach jej funkcjonowania, w tym w sferze bezpieczeństwa narodowego.

[242] Link prowadzący do aplikacji Face Swap Video by Deep Fake bezpośrednio w sklepie Google Play, https://play.google.com/store/apps/details?id=app.deepfaker.face_swap.ai_video_editor.gender_magic_face_merge_morph&hl=en_US [dostęp: 01.01.2023].

[243] Artykuł Roger Montti dotyczący możliwości dalszego rozwoju GPT-4, https://www.searchenginejournal.com/openai-gpt-4/476759/#close [dostęp: 21.01.2023].

[244] Artykuł Alana D. Thompson, komentujący wiadomości dotyczące kolejnych wersji czatu GPT-3, https://lifearchitect.ai/gpt-4/ [dostęp: 15.01.2023].

Rozdział 4. Percepcja zmanipulowanych przekazów wizualnych w świetle procesów psychologicznych i społecznych

Bezpieczeństwo narodowe odnosi się do ochrony interesów, integralności i suwerenności państwa oraz jego obywateli przed zagrożeniami. Jako szeroko definiowane pojęcie, zawsze jako podmiot ochrony stawiany jest naród, czyli obywatele danego państwa. Bezpieczeństwo narodowe obejmuje różne dziedziny, takie jak dyplomacja, polityka wewnętrzna czy zagraniczna, bezpieczeństwo energetyczne czy informacyjne, jak również środowiskowe. Wraz z postępem i globalizacją zwiększyła się liczba wektorów potencjalnych niebezpieczeństw, a bezpieczeństwo narodowe stało się jeszcze bardziej złożone i zróżnicowane. Tym samym wymaga koordynacji działań różnych służb i sektorów oraz współpracy międzynarodowej.

Jednym ze współczesnych zagrożeń dla bezpieczeństwa narodowego Polski są wrogie oddziaływania w sferze informacyjnej[245]. Celowe szerzenie nieprawdy lub półprawdy, manipulowanie obrazem i dźwiękiem, nagłaśnianie wygodnych faktów i ukrywanie tych niesprzyjających autorowi, to tylko niektóre z przykładów wpływu na umysły ludzi.

W niniejszym rozdziale przeanalizowana została przestrzeń informacyjna jako współczesne pole bitwy o umysły ludzi, czyli wojna kognitywna. W badaniu zadano cztery pytania zamknięte oraz jedno otwarte, bezpośrednio odnoszące się do rozpoznawania fałszu przez respondentów oraz ich zdolności percepcji.

Rozdział czwarty zbudowany jest z ośmiu podrozdziałów i opisuje przebieg procesu badawczego, mającego udzielić odpowiedzi na pytania dotyczące odbioru prawdziwych i fałszywych nagrań przez społeczeństwo. Pytanie badawcze brzmi – czy internauci rozróżniają materiały multimedialne prawdziwe od fałszywych? Hipoteza zakłada, iż przeciętny obywatel nie jest w stanie rozróżnić fałszywego nagrania od prawdziwego. Jednak celem weryfikacji niniejszej hipotezy, konieczne było dokonanie eksperymentu oraz analiza wyników.

[245] P. Kmiecik, „Bezpieczeństwo informacyjne Rzeczypospolitej w dobie Fake News – przykłady wykorzystania mediów cyfrowych w szerzeniu dezinformacji", Bezpieczeństwo Obronność Socjologia, 11/12, 2019.

Pierwszy podrozdział dotyczy oceny naturalności wyglądu prezentowanych nagrań. Na podstawie analizy odpowiedzi na zadane pytanie, przebadane zostało czy respondenci różnie oceniali punktowo fałszywe jak i prawdziwe nagrania, zarówno popularnych osób jak i tych nieznanych. W przypadku wykrycia statystycznych różnic w odpowiedzi na to pytanie pomiędzy grupami, mogłoby to wskazywać, iż nagrania deepfake były przez część respondentów nierozpoznane i zidentyfikowane jako naturalne.

W kolejnym podrozdziale przeanalizowano odpowiedzi na pytanie dotyczące prawdziwości wyświetlanych nagrań. Odpowiedzi udzielone przez respondentów mogą pomóc wskazać, jaka część z nich rozpoznała fałszywe filmy. Zestawienie wyników ze zmienną nominalną (pytanie końcowe z ankiety „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?") pozwoli również zweryfikować prawdziwość udzielonych odpowiedzi.

Trzeci podrozdział skupia analizę odpowiedzi na fakultatywne, opisowe pytanie dotyczące wskazania nienaturalnych elementów wyglądu. Przenalizowane treści podzielone zostały względem zmiennej nominalnej dotyczącej deklaracji rozpoznania nagrań deepfake (pytanie końcowe z ankiety „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").

Czwarty podrozdział dotyczy zaufania do wizerunku prezentowanych na nagraniach osób. Czy osoby występujące na nagraniach deepfake będą cieszyły się większym, czy mniejszym zaufaniem względem osób prezentowanych na pozostałych filmach? Czy osoby, które rozpoznały fałszywe filmy będą je gorzej oceniać?

Podrozdział piąty opisuje analizę zaufania do wizerunku osoby występującej na danym nagraniu. Omawiane w podrozdziale odpowiedzi na drugie pytanie mogą pozwolić określić poziom zaufania do średnio znanych influencerów, a także postaci prezentowanych na nagraniach deepfake. Czy osoby, które nie rozpoznały manipulacji będą lepiej oceniać zaufanie do osób z filmów deepfake, niż inni respondenci względem średnio znanych influencerów?

Szósty podrozdział ma na celu opisanie analizy odpowiedzi na pytanie dotyczące kojarzenia osób występujących na nagraniach. Pytanie to ma na celu weryfikację rozpoznawalności osób występujących na nagraniach wśród respondentów oraz udzielić odpowiedzi na pytanie czy osoby, które nie rozpoznały manipulacji deepfake znały

influencerów pod których próbowano się podszyć? Czy raczej nie rozpoznali oni filmów deepfake przez to, iż nie znali osób im prezentowanych na przykładach?

Przedostatni podrozdział dotyczy zaufania do filmowego przekazu. Przeanalizowane odpowiedzi respondentów mogą przyczynić się do odpowiedzi na pytanie czy nawet w momencie, gdy osoba orientuje się, iż nagranie jest fałszywe, nadal lepiej ocenia taki film niż nagranie prezentujące nieznane osoby. Ponadto zastanawiające jest czy zmiana twarzy na danym nagraniu może wpłynąć na ocenę filmowego przekazu.

Ostatni rozdział zawiera w sobie jednolity opis wyciągniętych na podstawie całego czwartego rozdziału wniosków. Zebrane z ośmiu podrozdziałów, udzielają odpowiedzi na stawiane w pracy trzecie pytanie pomocnicze, jednocześnie pozwalając zweryfikować stawianą w rozdziale hipotezę szczegółową.

4.1 Wygląd naturalny nagrań deepfake

W poniższym podrozdziale opisany został proces analityczny aspektów związanych z naturalnością prezentowanych nagrań, a ich odbiorem przez osoby oglądające je. Zastanawiające jest, czy osobom badanym udało się rozpoznać nagrania fałszywe od tych prawdziwych oraz w jakim stopniu były one przekonane o słuszności swojego wyboru.

Pytanie pierwsze, jakie zadano osobom biorącym udział w badaniu, po każdorazowym wyświetleniu nagrania brzmiało: na ile wyświetlony film wyglądał dla Ciebie naturalnie? Odpowiedź udzielano na skali 1 do 10, gdzie 1 to w ogóle, a 10 bardzo.

Filmy wyświetlano w losowej kolejności. Nagranie pierwsze oraz szóste były filmami deepfake, gdzie na twarz losowych osób nałożono twarze influencerów, przy czym wideo pierwsze podszywało się pod bardziej rozpoznawalną osobę niż nagranie szóste (pierwsze – ponad dwa miliony osób obserwujących na Instagramie, szóste milion). Nagrania drugie i trzecie, to filmy prezentujące nieznane, obce osoby, zachęcające do inwestycji. Filmy czwarty i piąty to nagrania średnio znanych influencerów, również zachęcających do tej samej inwestycji. Teksty wypowiadane na nagraniach były ze sobą zbieżne dla każdego filmu.

Zakładano, iż dla filmu pierwszego i ostatniego współczynnik naturalnego wyglądu będzie najmniejszy, zaś dla czwartego i piątego najwyższy. Przygotowywane

nagrania były bowiem celowo o obniżonej jakości, tak by łatwo można było zauważyć ich fałszywy charakter. Nagrania dwa i trzy, prezentujące losowe dwie osoby wyglądać mogły na nienaturalne, ze względu na brak medialnego doświadczenia osób na nich występujących. Tylko filmy czwarty i piąty nagrane były, za wynagrodzeniem, przez profesjonalnych influencerów, dzięki czemu powinny zostać najlepiej ocenione. Poniżej zaprezentowano statystki opisowe dla wszystkich nagrań dla odpowiedzi do pytania pierwszego.

Statystyki opisowe

	film1_01	**film2_01**	**film3_01**	**film4_01**	**film5_01**	**film6_01**
N	80	80	79	80	80	79
brakujące odpowiedzi	2	2	3	2	2	3
M	3.54	1.86	2.51	3.77	4.09	2.42
SE	0.278	0.192	0.254	0.290	0.287	0.220
95% CI dolna granica przedziału ufności dla średniej	2.99	1.49	2.01	3.21	3.53	1.99
95% CI górna granica przedziału ufności dla średniej	4.08	2.24	3.00	4.34	4.65	2.85
Me	3.00	1.00	1.00	3.50	4.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	2.49	1.72	2.26	2.59	2.57	1.95
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	10.0	10.0	10.0	10.0
SKE	0.675	2.09	1.84	0.490	0.274	1.63
SE_{SKE}	0.269	0.269	0.271	0.269	0.269	0.271
K	-0.704	3.43	3.15	-0.733	-0.896	2.63
SE_K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.875	0.574	0.710	0.881	0.907	0.755
p_{S-W}	< .001	< .001	< .001	< .001	< .001	< .001

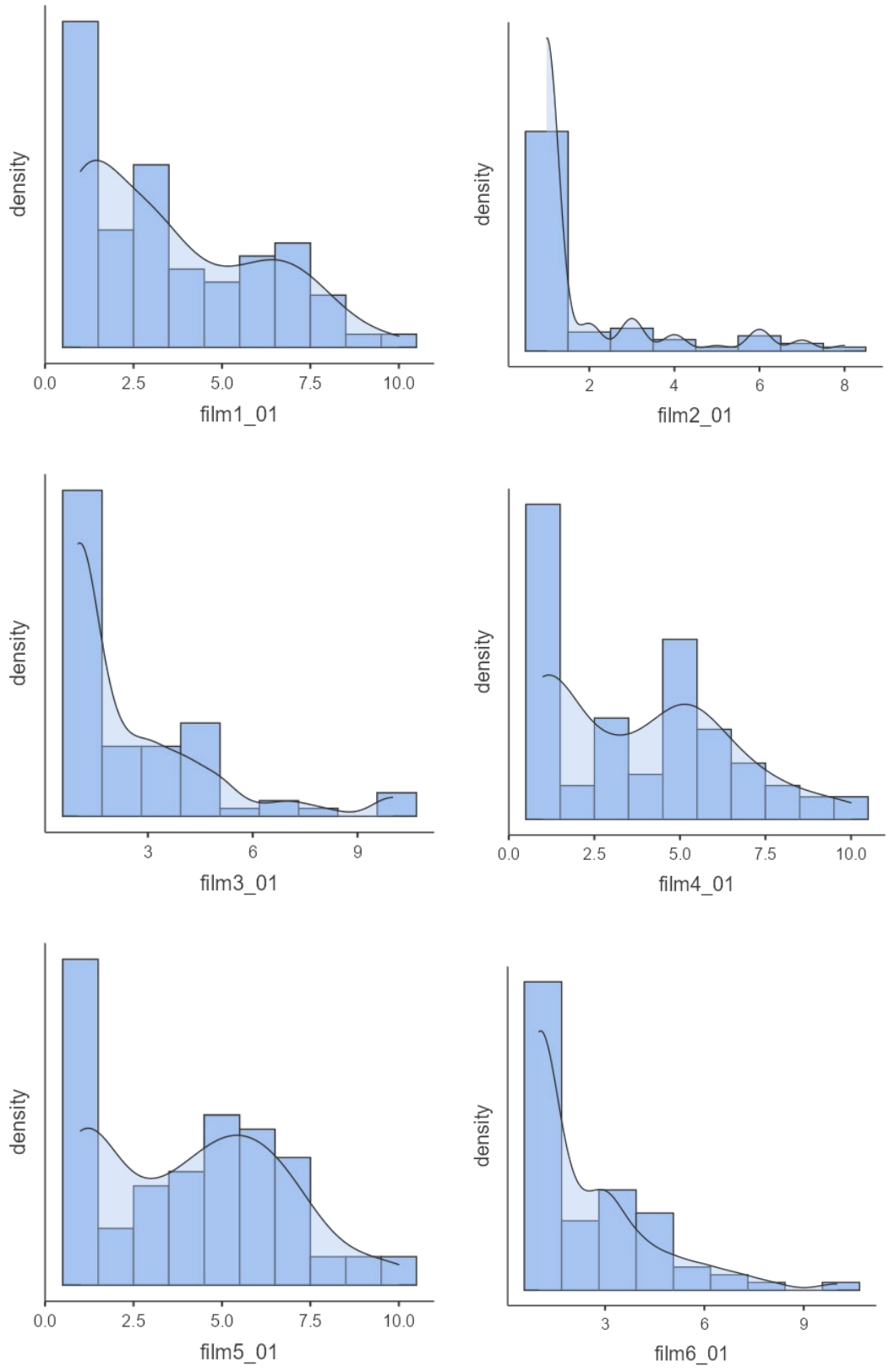
Tabela 1 Statystyki opisowe dla pierwszego pytania „Na ile wyświetlony film wyglądał dla Ciebie naturalnie?”.

Tabela 1 przedstawia statystyki opisowe dotyczące odpowiedzi na pytanie: „Na ile wyświetlony film wyglądał dla Ciebie naturalnie?”. W kolumnach zestawiono dane dotyczące odpowiedzi udzielonych po obejrzeniu każdego z sześciu filmów. Pierwsze nagranie prezentowało deepfake z udziałem znanej aktorki, promującej inwestycję na fałszywej platformie. Drugie i trzecie wideo przedstawiały nieznane osoby, również zachęcające do inwestycji. Nagrania czwarte i piąte ukazywały średnio popularnych influencerów promujących tę samą ofertę. Szóste wideo, podobnie jak pierwsze, było deepfake'em, tym razem z udziałem influencera o średniej rozpoznawalności, reklamującego oszukańczą inwestycję.

Dane z powyższej tabeli jednoznacznie pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników (dla odpowiedzi na pierwsze pytanie – w przypadku każdego filmu) jest zbliżony do rozkładu normalnego. Świadczą o tym – po pierwsze – wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$ (najwyższy raportowany – w większości nauk społecznych – poziom istotności statystycznej); po drugie – wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności (jeden ze stosowanych wyznaczników braku rozkładu normalnego), a w większości przypadków są one większe od 1 (kolejny ze stosowanych wskaźników braku rozkładu normalnego). Po trzecie – wartości bezwzględne kurtozy – które w każdym przypadku są większe od wartości błędu standardowego tej miary (jeden ze stosowanych wskaźników braku rozkładu normalnego), a w połowie przypadków – wartości kurtozy są większe niż 1 (kolejny ze stosowanych wskaźników braku rozkładu normalnego).

Ze względu na brak rozkładu normalnego uzyskanych wyników, w dalszych analizach statystycznych zastosowano testy nieparametryczne. Testy parametryczne, oparte na średniej, mogą być stosowane wyłącznie w przypadku rozkładu zbliżonego do normalnego. Ich użycie przy braku takiego rozkładu może prowadzić do błędnych wyników i nieprawidłowych wniosków.

Na poniższych wykresach przedstawiono histogramy uzyskanych wyników dla każdego z filmów w odniesieniu do pytania: „na ile wyświetlony film wyglądał dla Ciebie naturalnie?”. Histogramy te potwierdzają, że żaden z rozkładów wyników nie jest zbliżony do normalnego.



Wykres 2 Zbiór 6 wykresów odpowiedzi na pierwsze pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Dzieje

się tak zwłaszcza dla filmu 2, 3 i 6, gdzie tylko film 6 był przerobiony (deepfake), filmy 2 i 3 prawdziwe. Nagranie pierwsze (deepfake) jest pozytywnie oceniane przez prawie połowę osób je oceniających, co sugerować może, iż osoby te nie rozpoznały, iż jest to fałszywe nagranie.

Aby zweryfikować, czy odpowiedzi na pytanie dotyczące naturalności nagrania różniły się w zależności od tego, czy respondent rozpoznał deepfake, przeprowadzono dalszą analizę z uwzględnieniem zmiennej nominalnej E. Zmienna ta odnosiła się do pytania: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Celem było sprawdzenie czy osoby, które rozpoznały fałszywe nagrania, zrobiły to świadomie, czy reagowały w podobny sposób na wszystkie nagrania.

Poniższa tabela przedstawia analogiczne dane jak poprzednia tabela, z uwzględnieniem wspomnianej zmiennej nominalnej. Podobnie jak w tabeli nr. 1, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości skośności, gdzie w większości analizowanych przypadków są większe od wartości błędu standardowego skośności lub o zbliżonej wartości. Kolejnym elementem są wartości bezwzględne kurtozy, gdzie poza filmem 4 i 5 wartości te są większe niż wartości z błędu standardowego tej miary. Świadczą o tym również niskie wyniki testu Shapiro-Wilk (w większości przypadków $p < 0{,}001$, a w każdym $p < 0{,}05$).

Statystyki opisowe – Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?

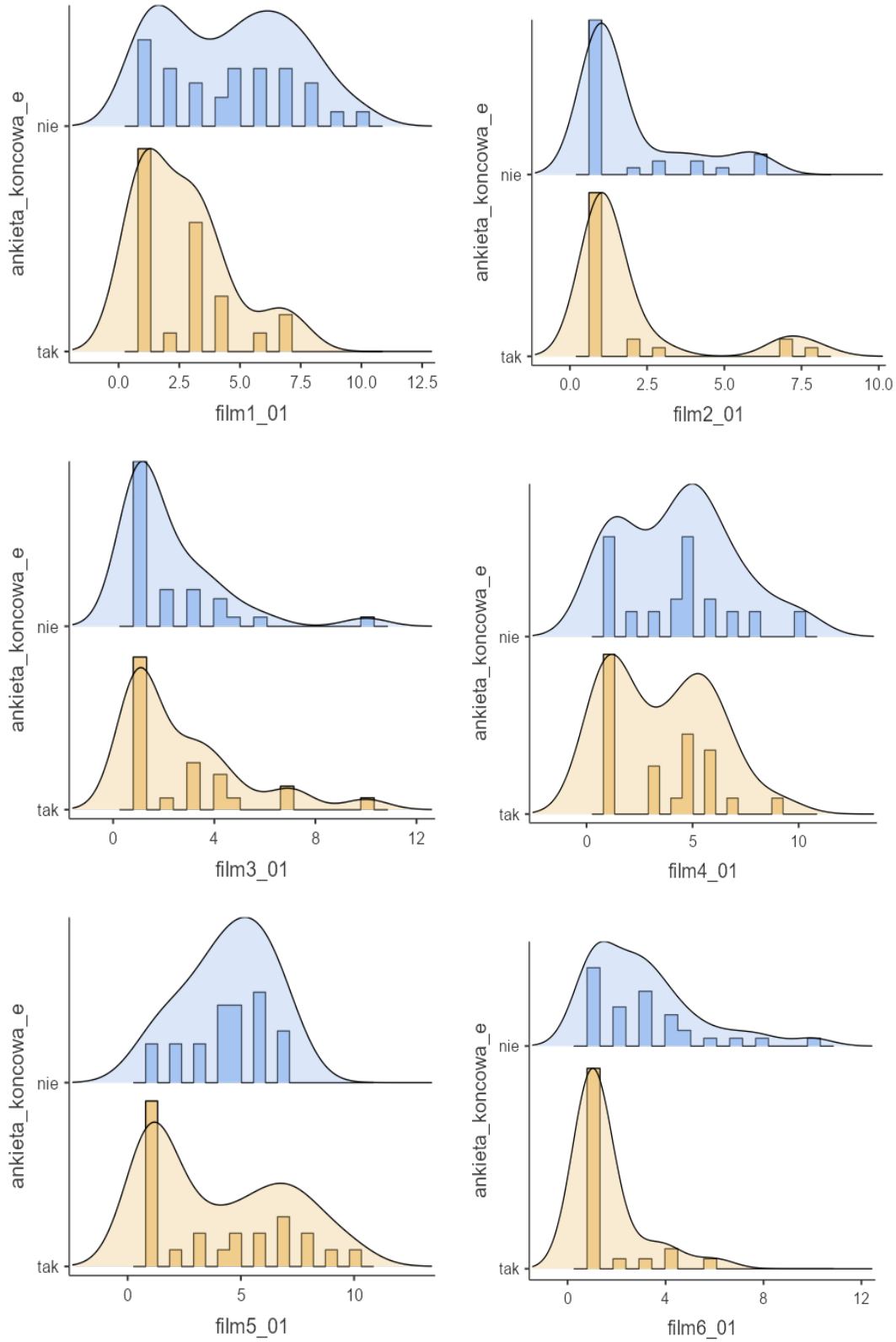
	zmienna nominalna E	**film1_01**	**film2_01**	**film3_01**	**film4_01**	**film5_01**	**film6_01**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	4.56	1.94	2.22	4.31	4.44	3.06
	tak	2.64	1.92	2.72	3.52	4.00	1.56
SE	nie	0.479	0.301	0.350	0.472	0.327	0.396
	tak	0.378	0.420	0.481	0.487	0.614	0.259
95% CI dolna granica przedziału ufności dla średniej	nie	3.62	1.35	1.53	3.39	3.80	2.29
	tak	1.90	1.10	1.78	2.57	2.80	1.05

Statystyki opisowe – Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?

	zmienna nominalna E	film1_01	film2_01	film3_01	film4_01	film5_01	film6_01
95% CI górna granica przedziału ufności dla średniej	nie	5.50	2.53	2.90	5.24	5.08	3.84
	tak	3.38	2.74	3.66	4.47	5.20	2.07
Me	nie	5.00	1.00	1.00	5.00	5.00	3.00
	tak	3.00	1.00	1.00	3.00	3.00	1.00
D	nie	1.00	1.00	1.00	1.00	6.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.71	1.70	1.98	2.67	1.85	2.24
	tak	1.89	2.10	2.41	2.43	3.07	1.29
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	6.00	10.0	10.0	7.00	10.0
	tak	7.00	8.00	10.0	9.00	10.0	6.00
SKE	nie	0.155	1.61	2.36	0.369	-0.434	1.42
	tak	1.09	2.32	1.61	0.388	0.451	2.43
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.14	1.17	6.89	-0.522	-0.787	2.02
	tak	0.527	4.05	2.37	-0.935	-1.28	5.49
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.930	0.603	0.674	0.915	0.925	0.838
	tak	0.803	0.497	0.751	0.857	0.850	0.508
PS-W	nie	0.039	< .001	< .001	0.016	0.028	< .001
	tak	< .001	< .001	< .001	0.002	0.002	< .001

Tabela 2 Statystyki opisowe dla pierwszego pytania „Na ile wyświetlony film wyglądał dla Ciebie naturalnie?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Poniżej przedstawiono histogramy dla każdego filmu. Analogicznie do powyższej tabeli, osoby badane zostały podzielone według zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").



Wykres 3 Zbiór 6 wykresów odpowiedzi na pierwsze pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na pierwsze pytanie, między osobami, które uważają, iż udało im się

rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 6 (deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe. Konieczna zdaje się być korelacja wyników odpowiedzi dla filmu pierwszego (deepfake) z filmem czwartym i piątym, gdyż ich układ zdaje się być podobny do siebie.

Aby sprawdzić, czy zaobserwowane różnice są istotne statystycznie, przeprowadzono serię testów Kruskala-Wallisa, będących nieparametrycznymi odpowiednikami jednoczynnikowej analizy wariancji (ANOVA). Wyniki tych testów zostały przedstawione w poniższej tabeli. Stwierdzono, że jedynie w przypadku pierwszego i ostatniego filmu (oba deepfake) rozkłady różniły się w sposób istotny statystycznie ($p < 0,05$). Oznacza to, że różnice między tymi dwoma grupami są na tyle znaczące, że można je wiarygodnie wyjaśnić na podstawie statystyk. Istotność statystyczna potwierdza, że różnice między grupami nie są przypadkowe i mogą być uznane za rzeczywiste w sensie statystycznym.

Test Kruskal-Wallis

	χ^2	df	p	ε^2
film1_01	7.3431	1	0.007	0.13113
film2_01	0.0786	1	0.779	0.00140
film3_01	0.4593	1	0.498	0.00820
film4_01	1.0934	1	0.296	0.01952
film5_01	0.6751	1	0.411	0.01206
film6_01	11.2844	1	< .001	0.20151

Tabela 3 Test Kruskal-Wallis dla odpowiedzi do pytania pierwszego.

Z powyższej tabeli wnioskować można, iż respondenci deklarujący rozpoznanie fałszywych nagrań odpowiadali na to pytanie w odmienny sposób niż osoby, które fałszu nie rozpoznały. Istotność statystyczna stwierdzona w przypadku filmu pierwszego i szóstego wskazuje, iż prawidłowo rozpoznano je jako fałszywe.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie pierwsze („Na ile wyświetlony film wyglądał dla Ciebie naturalnie?") będzie się różniła w przypadku różnych filmów, przeprowadzono nieparametryczny odpowiednik analizy wariancji z powtarzalnymi pomiarami (RM ANOVA), czyli test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 77,6; $p < 0,001$).

Aby zbadać różnice pomiędzy poszczególnymi odpowiedziami na pytanie pierwsze, przeprowadzono nieparametryczny odpowiednik testów post hoc. Zdecydowano się na porównania parami pomiędzy odpowiedziami dotyczącymi

wszystkich filmów, wykorzystując test Durbin-Conover. Wyniki tego testu zostały przedstawione w poniższej tabeli.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_01	-	film2_01	6.5799	< .001
film1_01	-	film3_01	4.5220	< .001
film1_01	-	film4_01	0.2979	0.766
film1_01	-	film5_01	0.9477	0.344
film1_01	-	film6_01	4.4949	< .001
film2_01	-	film3_01	2.0579	0.040
film2_01	-	film4_01	6.2821	< .001
film2_01	-	film5_01	7.5276	< .001
film2_01	-	film6_01	2.0850	0.038
film3_01	-	film4_01	4.2241	< .001
film3_01	-	film5_01	5.4697	< .001
film3_01	-	film6_01	0.0271	0.978
film4_01	-	film5_01	1.2456	0.214
film4_01	-	film6_01	4.1971	< .001
film5_01	-	film6_01	5.4426	< .001

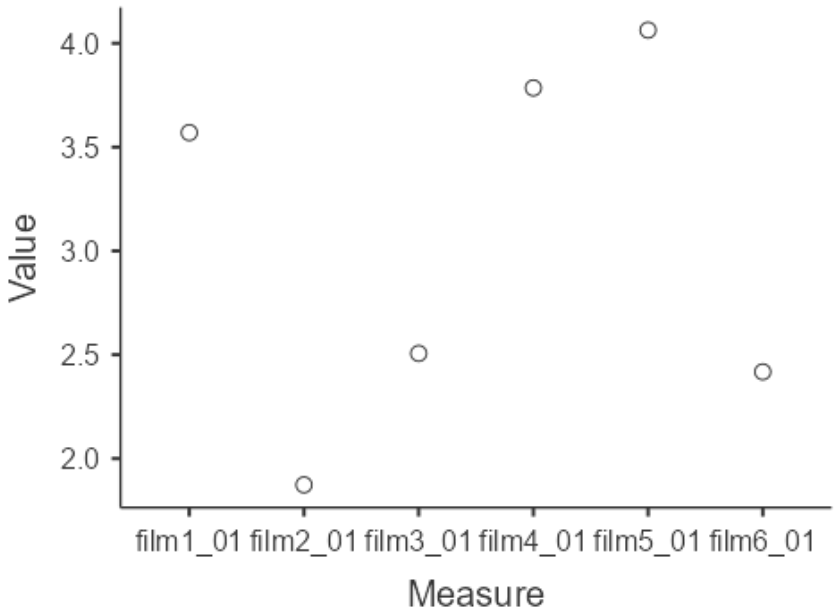
Tabela 4 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania pierwszego.

Powyższe obliczenia pokazały, że film 1 (deepfake) różni się od filmu 2, 3 i 6, na co wskazuje $p < 0,05$, natomiast nie zaobserwowano różnicy względem filmów 4 i 5 ($p > 0,05$). Potwierdza to wcześniejsze założenia, iż film wykonany z użyciem technologii deepfake, prezentujący znaną osobę, jest podobnie odbierany pod względem naturalności jak pozostałe nagrania (prawdziwe) również wykorzystujące wizerunki znanych osób. Nagrania 2, a 3 również nie są istotnie różne od siebie (oba prezentują nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4 oraz 5. Zastanawiająca jest różnica pomiędzy nagraniem 3, a 6 ($p < 0,05$), której nie zaobserwowano w zestawieniu nagrań 2, a 6 (nie zaobserwowano różnicy). Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również nie zaobserwowano różnicy. Film 4 i 5 są natomiast różne od filmu 6 (deepfake, $p < 0,05$).

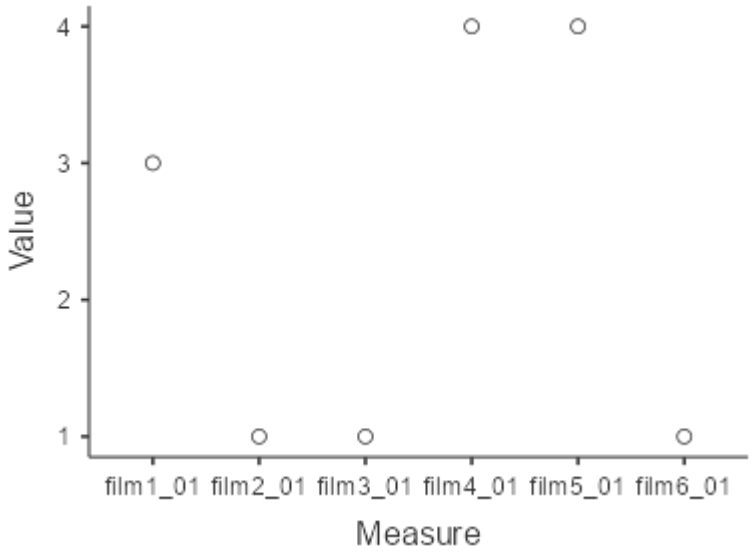
Wnioskować można, iż film 1 nie był istotnie różny od filmów 4 i 5, co świadczy o ich podobnym odbiorze przez widzów. Celem weryfikacji tej hipotezy, poniżej umieszczono również wykresy przedstawiające średnie oraz wykresy przedstawiające mediany. Wynika to z tego, iż rozkłady wyników są znacząco różne od rozkładu

normalnego. Średnie są więc w tym przypadku mocno obciążoną miarą tendencji centralnej, w związku z czym zaleca się stosowanie mediany.

Mediana pokazuje, iż filmy 2 oraz 3 i 6 nie wzbudziły zaufania wśród osób je oglądających. Nagrania 1 (nieprawdziwe), 4 i 5 są jednak zauważalnie blisko oceniane, co widać zarówno po średniej jak i medianie.



Wykres 4 Średnia odpowiedzi dla pytania pierwszego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 5 Mediana odpowiedzi dla pytania pierwszego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Obserwując rozkład median i średnich na wykresie, zauważyć można, iż film pierwszy (deepfake) jest podobnie oceniany do filmów ze znanymi osobami (4 i 5).

Natomiast nagrania 2 i 3, prezentujące nieznane osoby jak i film 6 (deepfake, influencer) są istotnie gorzej oceniane (M = 1).

Pierwsze z nagrań deepfake zostało pod względem naturalności podobnie odebrane jak pozostałe, prawdziwe filmy znanych osób. Słusznie założono, iż współczynnik naturalnego wyglądu największy będzie dla prawdziwych nagrań 4 i 5, natomiast zaskakujący jest wynik filmu 1, który oceniony został do nich znacząco podobnie. Co więcej, pomimo istotności statystycznej różnic, osoby, które oświadczyły, iż rozpoznały fałsz, często również oceniały go jako wyglądający naturalnie. Tylko w zakresie filmu szóstego, zauważono zdecydowaną odmianę punktowania przez respondentów wyglądu naturalnego między grupami.

Najniżej naturalny wygląd oceniony został w filmach prezentujących nieznane osoby – nagrania drugie i trzecie. Z porównania Durbin-Conover wiadomo, iż nie zaobserwowano różnicy pomiędzy filmem drugim, a szóstym (deepfake). Wszystkie te trzy można więc wstępnie zakwalifikować jako jednakowo nienaturalne, sprawiające wrażenie nieprawdziwych.

4.2 Weryfikacja prawdziwości nagrań

W poniższym podrozdziale przenalizowane zostały aspekty związane z prawdziwością prezentowanych nagrań oraz rozpoznawaniem fałszywych nagrań przez osoby oglądające je. Zastanawiające jest czy osobom badanym udało się rozpoznać obrazy fałszywe od tych prawdziwych oraz w jakim stopniu były one przekonane o słuszności swojego wyboru. Zastanawia również, czy osoby, które twierdzą, iż rozpoznały fałszywe nagrania, faktycznie prawidłowo go zidentyfikowały.

Pytanie, jakie zadano osobom biorącym udział w badaniu, po każdorazowym wyświetleniu nagrania brzmiało: na ile przekonuje Cię prawdziwość powyższego nagrania? Odpowiedź udzielano na dziesięciostopniowej skali, gdzie 1 to w ogóle, 10 bardzo.

Założono, iż dla filmu szóstego (deepfake) poziom rozpoznania deepfake będzie większy niż dla pierwszego (fałszywego) nagrania. Wynikać może to z gorszego przygotowania nagrania oraz słabszej gry aktorskiej osoby oryginalnie w nim występującej. Wpływać na to może również słabsza rozpoznawalność zamieszczonego tam influencera. Film pierwszy może jednak osiągnąć zbliżony wynik do nagrań 4 i 5. Film czwarty i piąty jednak powinny otrzymać najwyższe oceny, ze względu na ich

naturalną prawdziwość oraz dobrze opanowaną grę aktorską osób w nich występujących. Nagrania drugie i trzecie, prezentujące nieznane osoby, podobnie jak filmy deepfake, wyglądać mogły na sztuczne, ze względu na brak medialnego doświadczenia osób występujących oraz kiepską jakość nagrań. Poniżej zaprezentowano statystki opisowe dla wszystkich filmów.

Statystyki opisowe

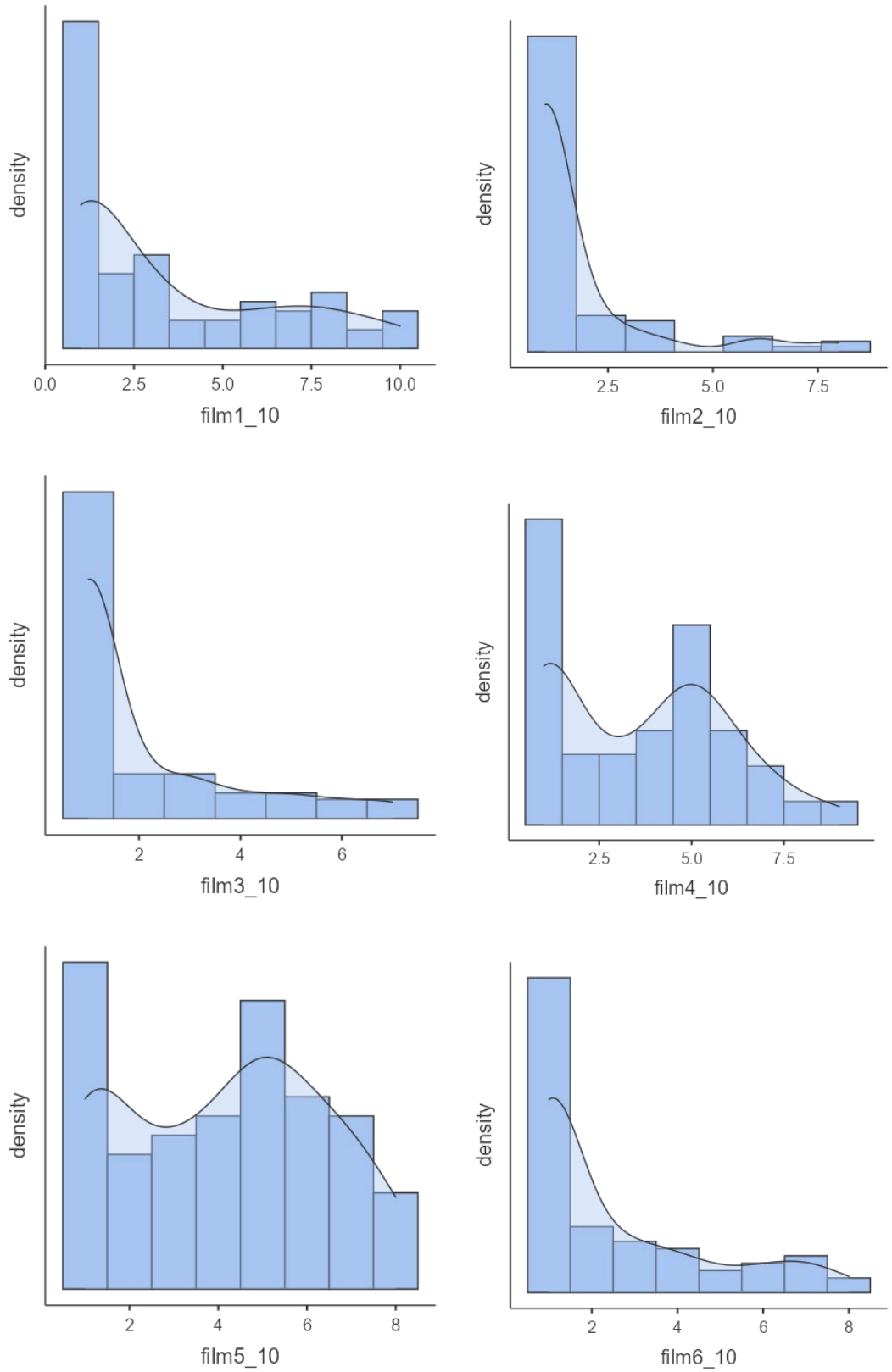
	film1_10	**film2_10**	**film3_10**	**film4_10**	**film5_10**	**film6_10**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	3.40	1.70	2.04	3.63	4.11	2.48
SE	0.327	0.183	0.194	0.260	0.253	0.236
95% CI dolna granica przedziału ufności dla średniej	2.76	1.34	1.66	3.11	3.62	2.02
95% CI górna granica przedziału ufności dla średniej	4.04	2.06	2.42	4.14	4.61	2.94
Me	2.00	1.00	1.00	4.00	4.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	2.93	1.63	1.73	2.33	2.26	2.10
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	7.00	9.00	8.00	8.00
SKE	0.965	2.68	1.62	0.333	-0.0016	1.28
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-0.449	6.56	1.53	-0.964	-1.21	0.385
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.791	0.496	0.659	0.887	0.917	0.733
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 5 Statystyki opisowe dla dziesiątego pytania „Na ile przekonuje Cię prawdziwość powyższego nagrania?”.

Powyższa tabela przedstawia statystki opisowe dotyczące odpowiedzi na dziesiąte pytanie: „na ile przekonuje Cię prawdziwość powyższego nagrania?”. w kolumnach znajdują się statystyki dotyczące odpowiedzi na to pytanie po obejrzeniu każdego z sześciu filmów. Dane z powyższej tabeli jednoznacznie pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników (dla odpowiedzi na pytanie dziesiąte – w przypadku każdego filmu) jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$, wartości skośności, które w przypadku większości nagrań, poza filmem piątym, są większe od wartości błędu standardowego skośności, a co więcej, w większości przypadków, są one większe od 1 (oprócz filmu 4 i 5). Świadczą o tym również wartości bezwzględne kurtozy, które w większości przypadków (poza filmami

deepfake – 1 i 6) są większe od wartości błędu standardowego tej miary, a w połowie przypadków wartości kurtozy są większe niż 1. W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Na poniższych wykresach znajdują się histogramy otrzymanych wyników – dla każdego filmu osobno. Pytanie dziesiąte brzmiało: „Na ile wyświetlony film wyglądał dla Ciebie naturalnie?”. Histogramy te również pokazują, iż rozkład wyników dla żadnego z filmów nie jest zbliżony do rozkładu normalnego, a każdy z wykresów cechuje silna prawoskośność.



Wykres 6 Zbiór 6 wykresów odpowiedzi na dziesiąte pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi, zwłaszcza dla filmów 1, 2, 3 i 4. Wyłącznie ocena filmu 5 ocena zdaje się być równomiernie rozłożona. Jednak

pomimo dużej liczby odpowiedzi „5” (podobnie jak w czwartym filmie), dominującą jest 1 punkt. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – w ogóle. Dzieje się tak zwłaszcza dla filmu 2, 3 oraz częściowo 6, gdzie tylko film 6 był przerobiony (deepfake), a filmy 2 i 3 prawdziwe.

Poniższa tabela przedstawia analogiczne dane jak poprzednia tabela, natomiast w tym przypadku rozkłady zmiennych zostały podzielone względem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Konieczne było bowiem sprawdzenie czy osoby, które rozpoznały fałsz, uczyniły to świadomie czy reagowały jednakowo na każdy inne film. Zastanawiające jest również, czy osoby te uznały nagrania 2 i 3 również jako filmy deepfake.

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to między innymi wartości skośności, które dla większości nagrań (poza filmem 4 i 5) są większe od bezwzględnej wartości błędu standardowego skośności. Świadczy o tym również wartość bezwzględna kurtozy, która poza grupą osób udzielających odpowiedzi „nie” względem zmiennej nominalnej w nagraniu 4 i 5, jest wyższa od wartości błędu standardowego. Świadczą o tym również niskie wyniki ($p < 0,05$) testu Shapiro-Wilk dla każdego z nagrań (a w większości $p < 0,001$).

Statystyki opisowe

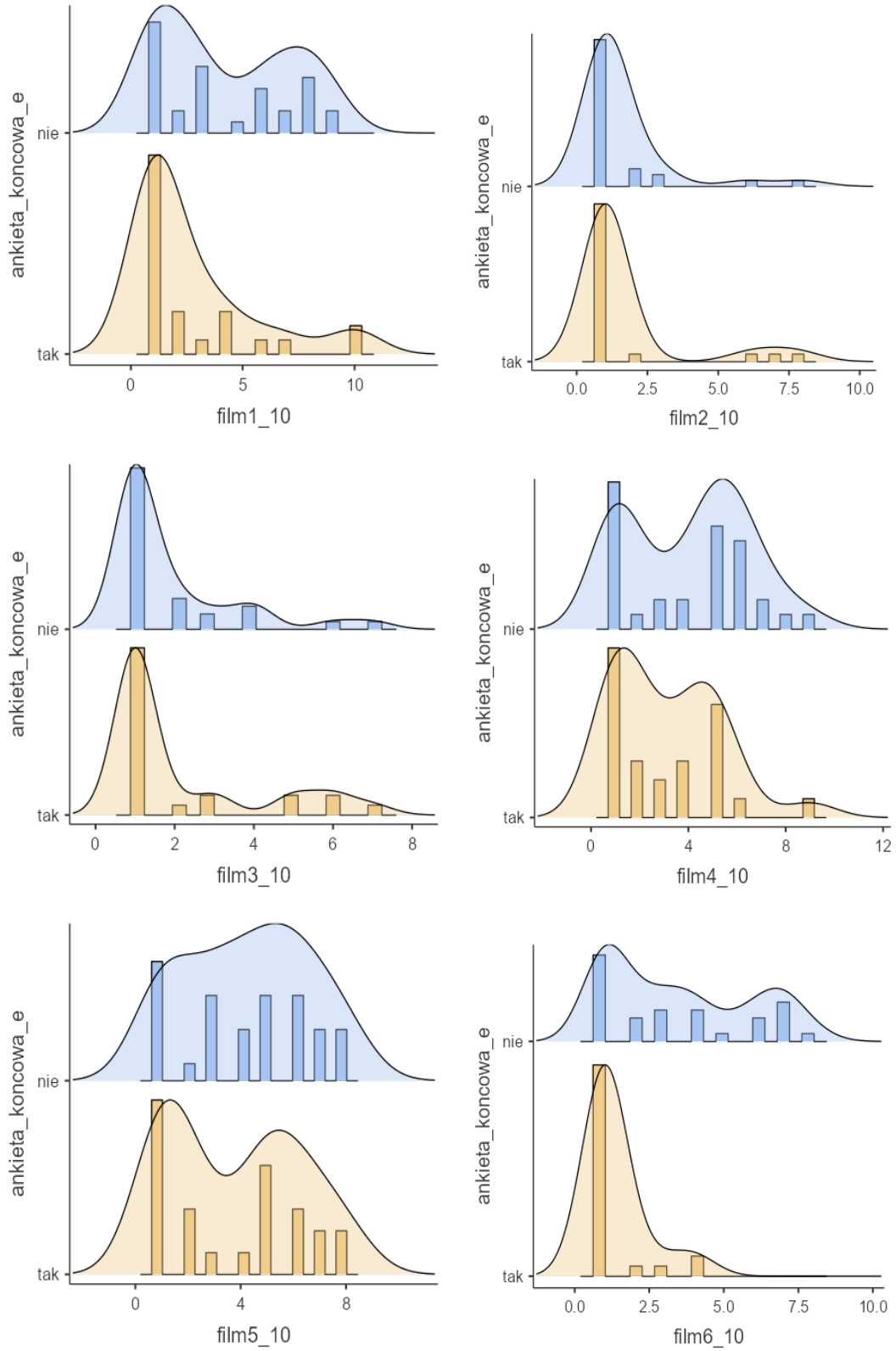
	zmienna nominalna E	**film1_10**	**film2_10**	**film3_10**	**film4_10**	**film5_10**	**film6_10**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	4.16	1.59	1.88	4.00	4.25	3.47
	tak	2.72	1.76	2.16	3.12	3.76	1.36
SE	nie	0.520	0.273	0.276	0.433	0.414	0.426
	tak	0.552	0.401	0.394	0.429	0.501	0.181
95% CI dolna granica przedziału ufności dla średniej	nie	3.14	1.06	1.33	3.15	3.44	2.63
	tak	1.64	0.974	1.39	2.28	2.78	1.00

Statystyki opisowe

	zmienna nominalna E	film1_10	film2_10	film3_10	film4_10	film5_10	film6_10
95% CI górna granica przedziału ufności dla średniej	nie	5.18	2.13	2.42	4.85	5.06	4.30
	tak	3.80	2.55	2.93	3.96	4.74	1.72
Me	nie	3.00	1.00	1.00	5.00	4.50	3.00
	tak	1.00	1.00	1.00	3.00	4.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.94	1.54	1.56	2.45	2.34	2.41
	tak	2.76	2.01	1.97	2.15	2.50	0.907
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	9.00	8.00	7.00	9.00	8.00	8.00
	tak	10.0	8.00	7.00	9.00	8.00	4.00
SKE	nie	0.323	3.27	2.01	0.0421	-0.0322	0.469
	tak	1.76	2.55	1.46	0.820	0.227	2.45
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.53	11.0	3.66	-1.17	-1.19	-1.28
	tak	2.29	5.18	0.637	0.432	-1.45	4.87
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.850	0.452	0.636	0.884	0.918	0.854
	tak	0.681	0.430	0.641	0.854	0.867	0.450
PS-W	nie	< .001	< .001	< .001	0.003	0.018	< .001
	tak	< .001	< .001	< .001	0.002	0.004	< .001

Tabela 6 Statystyki opisowe dla dziesiątego pytania „Na ile przekonuje Cię prawdziwość powyższego nagrania?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Poniżej przedstawiono histogramy dla każdego filmu, przy czym badani zostali podzieleni zgodnie ze zmienną nominalną E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").



Wykres 7 Zbiór 6 wykresów odpowiedzi na dziesiąte pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na dziesiąte pytanie, między osobami, które uważają, iż udało im się

rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 6 (deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe.

W celu oceny, czy zaobserwowane różnice są statystycznie istotne, przeprowadzono serię testów Kruskala-Wallisa. Wyniki tych analiz przedstawiono w poniższej tabeli. Stwierdzono, że jedynie dla pierwszego i ostatniego filmu (oba deepfake) różnice w rozkładach są istotne statystycznie ($p < 0,05$). Oznacza to, że różnice między tymi grupami są wystarczająco duże, aby można je było wiarygodnie wyjaśnić statystycznie. Statystyczna istotność sugeruje, że różnice te nie są przypadkowe i mogą być uznane za rzeczywiste.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_10	3.9748	1	0.046	0.07098
film2_10	0.1768	1	0.674	0.00316
film3_10	0.0133	1	0.908	2.37e-4
film4_10	2.2095	1	0.137	0.03946
film5_10	0.5516	1	0.458	0.00985
film6_10	14.7096	1	< .001	0.26267

Tabela 7 Test Kruskal-Wallis dla odpowiedzi do pytania dziesiątego.

Z powyższej tabeli wnioskować można, iż respondenci deklarujący rozpoznanie fałszywych nagrań odpowiadali na to pytanie w odmienny sposób niż osoby, które fałszu nie rozpoznały. Istotność statystyczna stwierdzona w przypadku filmu pierwszego, ale przede wszystkim szóstego, świadczyć może, iż duża liczba respondentów prawidłowo rozpoznała je jako fałszywe.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie dziesiąte („Na ile przekonuje Cię prawdziwość powyższego nagrania?") będzie różniła się w przypadku różnych filmów, przeprowadzono test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 111; $p < 0,001$).

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na pytanie dziesiąte, przeprowadzono test Durbin-Conover. Tabelę z wynikami testu zamieszczono poniżej.

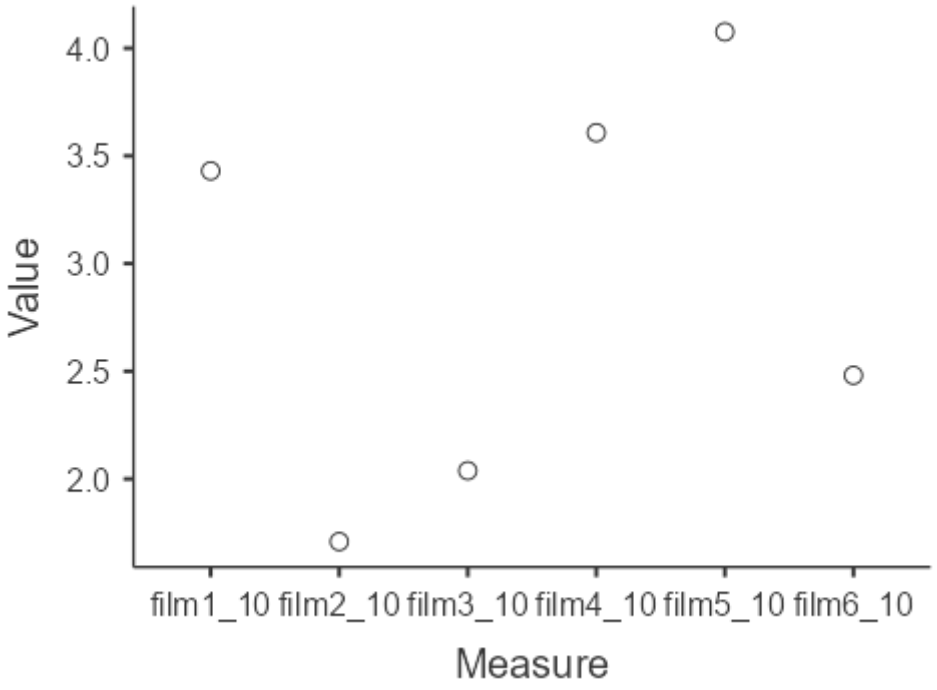
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_10	-	film2_10	5.94	< .001
film1_10	-	film3_10	4.80	< .001
film1_10	-	film4_10	1.80	0.073
film1_10	-	film5_10	4.06	< .001
film1_10	-	film6_10	2.88	0.004
film2_10	-	film3_10	1.15	0.252
film2_10	-	film4_10	7.74	< .001
film2_10	-	film5_10	10.01	< .001
film2_10	-	film6_10	3.06	0.002
film3_10	-	film4_10	6.59	< .001
film3_10	-	film5_10	8.86	< .001
film3_10	-	film6_10	1.91	0.057
film4_10	-	film5_10	2.27	0.024
film4_10	-	film6_10	4.68	< .001
film5_10	-	film6_10	6.94	< .001

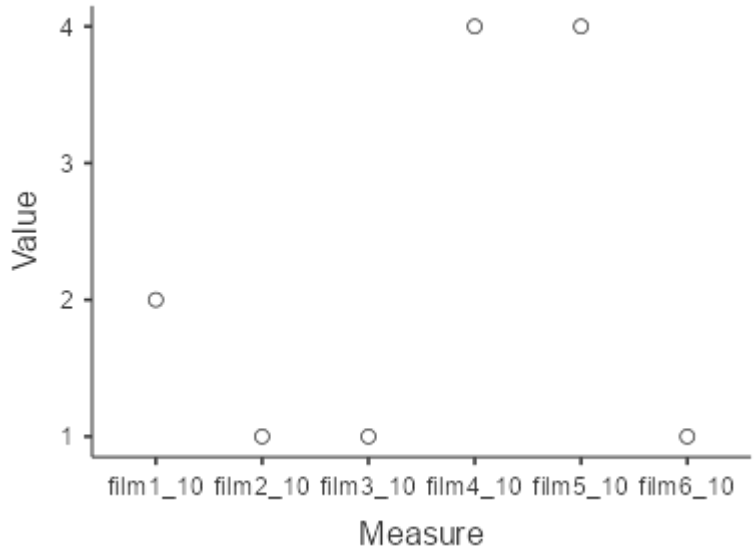
Tabela 8 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania dziesiątego.

Powyższe obliczenia pokazały, że film 1 (deepfake) różni się od filmu 2, 3, 5 i 6, na co wskazuje $p < 0,05$, natomiast nie zaobserwowano różnicy pomiędzy nim a filmem 4 ($p > 0,05$). Potwierdza to wcześniejsze założenia, iż dobrze wykonany film deepfake, prezentujący znaną osobę, może być trudno rozpoznawalny. Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4 oraz 5. Nagranie 2 różni się również z nagraniem 6 (deepfake), jednak nie zaobserwowano różnicy między filmem 3, a 6 ($p > 0,05$). Zastanawiająca jest różnica pomiędzy nagraniem 3, a 6 ($p < 0,05$), której nie zaobserwowano w zestawieniu nagrań 2, a 6 (nie zaobserwowano różnicy). Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również nie zaobserwowano różnicy. Film 4 i 5 są natomiast różne od filmu 6.

Wnioskować można, iż film 1 nie był istotnie różny od filmu 4, a film 6 nie był istotnie różny od filmu 3. Świadczyć to może o ich podobnym odbiorze pod względem prawdziwości przez respondentów. Celem weryfikacji tej hipotezy, poniżej umieszczono wykresy przedstawiające średnie oraz wykresy przedstawiające mediany. Wynika to z tego, iż rozkłady wyników są znacząco różne od rozkładu normalnego.



Wykres 8 Średnia odpowiedzi dla pytania dziesiątego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 9 Mediana odpowiedzi dla pytania dziesiątego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Mediana pokazuje, iż filmy 2, 3 oraz 6 (deepfake) nie zostały zidentyfikowane jako prawdziwe wśród respondentów (M = 1). Biorąc pod uwagę wyniki Test Kruskal-Wallis należy jednak stwierdzić, iż tylko w przypadku nagrania 6 różnice oceny między grupami względem zmiennej nominalnej E były istotnie różne. Sugeruje to, iż powód nieprawdziwości nagrań 2 i 3 był odmienny od tego, który zdefiniowany został

w zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”).

Nagranie 1 (deepfake) oraz 4 i 5 (prawdziwe) są zauważalnie blisko oceniane, co widać głównie po średniej, ale również medianie. Zaświadcza o tym również wynik testu Durbin-Conover, wskazujący, iż film 1 nie był istotnie różny w odbiorze prawdziwości od filmu 4. Potwierdza to wstępne założenia, iż film 6 będzie rozpoznawany w większej liczbie przypadków jako deepfake, niż film 1, częściej identyfikowany jako prawdziwy. Przyczyny różnic wskazane zostały w podrozdziale 4.3.

Wnioskować można, iż duża część respondentów rozpoznała nagrania deepfake, jednak część z nich zidentyfikowała je wyłącznie na filmie 6. Nagranie 1 u części respondentów, twierdzących, iż rozpoznała deepfake, nie został przez nie rozpoznany. Świadczy o tym średnia zbliżona do wyników filmów 4 i 5, zwłaszcza biorąc pod uwagę grupę osób odpowiadających „Tak” dla pytania „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Dla tej grupy, dla filmu pierwszego średnia wyniosła 2,72 zaś dla filmu czwartego 3,12. Świadczyć o tym może również wynik testu Durbin-Conover, mówiący, iż film 1 różnił się od 6 ($p < 0,05$).

4.3 Pytanie opisowe – nienaturalne elementy

Analizując odbiór nagrań przez osoby biorące udział w badaniu, nie sposób pominąć ogólne odczucia respondentów im towarzyszące. Różnicę w percepcji nagrań starano się uchwycić poprzez zaproponowanie badanym osobom fakultatywnego pytania do opisowej odpowiedzi.

Po obejrzeniu kolejno każdego z filmów i odpowiedzi na 15 zamkniętych pytań, każdy z respondentów miał sposobność uszczegółowienia swoich odpowiedzi, wyrażenia uczuć oraz pozostawienia uwag w formie krótkiego opisu, odpowiadającego na dwa, otwarte pytania. Pierwsze z opisowych pytań brzmiało: „jakie elementy aktora / aktorki były dla ciebie nienaturalne (sztuczne) na powyższym nagraniu?”, zaś drugie, bardziej ogólne: „co sądzisz o tym nagraniu?”. W poniższych pod-podrozdziałach przenalizowano odpowiedzi na pierwsze z prezentowanych powyżej pytań. Miało ono na celu zebrać od respondentów ich subiektywne odczucia względem nienaturalnych elementów oraz informacje o tym, które składowe danego nagrania mogły świadczyć o jego podrobieniu, modyfikacji. Prezentację wyników dokonano względem zmiennej nominalnej E – „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania

przygotowane technologią deepfake?". W pierwszym podpodrozdziale umieszczono odpowiedzi osób, które twierdziły, iż nie rozpoznały deepfake, natomiast w dalszej części – tych respondentów, którzy twierdzili, iż rozpoznali nagrania deepfake. Pominięto opinie osób, które nie wiedziały czy rozpoznały nagrania deepfake czy nie.

Analizy odpowiedzi dokonano w podziale na grupy filmów: wpierw opisywane są odczucia względem pierwszego z filmów deepfake (film 1) następnie drugie z nagrań (film 6). W dalszej kolejności poddano zbiorowej analizie opisy dla filmów prezentujących nieznane osoby (filmy 2 i 3) oraz średnio znanych celebrytów (filmy 4 i 5)[246].

Celem niniejszego podrozdziału jest próba znalezienia elementów mogących świadczyć o niedopracowaniu nagrań deepfake, a tym samym ich rozpoznaniu. Ponadto zastanawiające są ogólne odczucia respondentów towarzyszące im przy oglądaniu nagrań. Zaznaczyć należy, iż cały czas byli oni przekonani, iż celem rzeczywistym doświadczenia jest badanie najskuteczniejszej formy reklamy marketingowej – nagrań wideo – dla usług finansowych oraz oszacowania wpływu popularności wizerunku osób na odbiór kampanii.

4.3.1 Nierozpoznanie nagrań deepfake – uwagi dotyczące nienaturalnych elementów

W części pierwszej poniższego pod-podrozdziału opisane zostały przykłady odpowiedzi na pierwsze pytanie, spośród osób które twierdzą, iż nie rozpoznały nagrań deepfake spośród prezentowanych filmów.

Analizując odpowiedzi respondentów na pytanie dotyczące nienaturalnych elementów w pierwszym filmie deepfake, można, iż zwracali oni uwagę przede wszystkim na nienaturalność głosu. „Denerwujący miauczący głos, sztywna postawa i wzrok wlepiony w rozmówcę", „Mówienie ściszonym głosem. Zastosowanie filtra podczas mówienia o współpracy.", „Denerwujący miauczący głos, sztywna postawa i wzrok wlepiony w rozmówcę", „niepewny ton głosu", „Ruchy ust", „Po raz kolejny to samo zdanie wykorzystane, a jest to inny filmik". Przytoczone odpowiedzi podkreślają nienaturalność w mowie, zarówno w kontekście tonu głosu, jak i wykorzystania filtra lub efektu dźwiękowego.

[246] Cytowane w niniejszym podrozdziale opinie zostały przytoczone w niezmienionej, pełnej formie. Uwzględnia to błędną ortografię, interpunkcję oraz zastosowane buźki (emotikony).

Ponadto respondenci zwracają uwagę na brak autentyczności influencerki. „zbyt smutny wyraz twarzy i mało energii", „To jest influencerka, większość jej filmików jest sztuczna". Brak autentyczności powtórzony został również w kontekście roli celebryckiej odgrywanej przez influencerka. „Ogólnie jest wiarygodna, ale nie ufam współpracom z celebrytami", „Nie mówiła przekonywująco", „Typowa reklama". W kontekście zarządzania finansami i inwestowaniem, respondenci wyrażają nieufność wobec współprac z celebrytami.

Analizując odpowiedzi respondentów na pytanie dotyczące nienaturalnych elementów w drugim filmie deepfake (nagranie 6), przede wszystkim należy zwrócić uwagę na dużą ilość odpowiedzi komentujących nienaturalność w mowie. „Mówienie z dużą przerwą pomiędzy kolejnymi zdaniami", „Przyspieszony sposób wypowiedzi, przerwy w wypowiedzi, niespokojny oddech i łapanie powietrza", „Momenty, w których bardzo przyspieszał wypowiedź. Znowu odniesienie do inflacji, dobór słów", „Przerysowane, niepotrzebny filtr, sztuczny ton głosu", „głos", „Nałożony filtr. Mówi jakby czytał z kartki". Respondenci zauważają nienaturalny ton głosu oraz przerwane, przyspieszone lub przerysowane wypowiedzi, co sugeruje, że influencer mógł udawać – czytać tekst z kartki lub korzystać ze specjalnych efektów dźwiękowych wzmacniający jego przekaz.

Ponadto respondenci ponownie zwracają uwagę na niezachowanie autentyczności, przez co cały film wydaje się być sztucznym. „Flirt na tworzy aktora? :/ facet chodzi na silownie testowac suplementy diety?! Gratuluje, prawdopodobnie testuje tez inwestowanie na swoich pieniądzach. Tak samo jak chodzenie na silownie jak i inwestowanie przez aktora jest jednym wielkim fejkiem", „Osoba ta jest mało przekonywująca", „Lokowanie nazwy firmy nigdy chyba nie brzmi naturalnie". Respondenci zauważają brak przekonującej wypowiedzi oraz uznają wizerunek influencera za niespójny z reklamowaną przez niego inwestycją.

Odpowiedzi respondentów dla filmów deepfake, w grupie, która twierdzi, iż nie rozpoznała tej technologii, skupiają się głównie wokół nienaturalności w mowie czy niespójności wizerunków influencerów z prezentowanym przez nich przekazem. Kilka z przytoczonych powyżej odpowiedzi potwierdza, iż faktycznie osoby oglądające nagrania nie rozpoznały ich fałszerstwa przy pomocy technologii deepfake.

Analizując odpowiedzi respondentów dla filmów osób nieznanych (nagrania 2 i 3), można wyróżnić kilka głównych obserwacji, z których najliczniejsza dotyczy braku

autentyczności w zachowaniu. „Wymuszony uśmiech, gestykulacja i ruchy ciała, czytanie tekstu, podnoszenie głosu na koniec niektórych fraz", „Pani wzrokiem jest gdzies indziej. Wyglada to tak jakby czytala z kartki obok, a nie mowila to co mysli", „Nienaturalne wydaje się to, że facet ktory proponuje coś co przynosi mu „„zyski"„ miałby chwalić sie tym oraz powodować wzrost konkurencyjnosci w swojej branży. Na ogół jest raczej odwrotnie.", „Mężczyzna prawdopodobnie mówił tekst, którego wcześniej się nauczył, zachowywał się sztucznie", „Wyuczone, wcześniej przygotowane zdania". Respondenci zauważają nienaturalne zachowania, gesty, mimikę twarzy oraz sposób mówienia, co sugeruje, że aktorzy mogą wydawać się mało autentyczni lub wyuczają swoje wypowiedzi. Ponadto kilkukrotnie pojawiają się uwagi dotyczące odczucia jakby osoby występujące na tych filmach czytały swoje wypowiedzi z kartki. „Widać, że musiała nagrać ten filmik patrzenie po bokach, nienaturalny ton mówienia, gest na końcu filmu", „Sztuczne, widać, że Pani czytała z kartki", „Wzrok idzie w bok, jakby czytała z kartki. Gra aktorska rodem z trunych spraw.", „Mówienie tak jakby wyuczonych na pamięć zdań.", „Kobieta czytała tekst prawdopodobnie z kartki, nienaturalnie się zachowywała", „Nie patrzyła w kamerę, czytała tekst wyświetlany obok", „Gość mówi jakby czytał z kartki.", „Mało emocji na twarzy. Tak jakby czytanie, tego co ma się do powiedzenia a nie mówienie własnymi słowami.", „Pan ewidentnie mówił czyjś wyuczony tekst, jakby ktoś mu kazał". Wiele odpowiedzi wskazuje na to, że odczucie jakby ktoś czytał tekst z kartki wzięło się przede wszystkim z nienaturalnego tonu wypowiedzi oraz patrzenia w bok zamiast w obiektyw kamery. Niektóre odpowiedzi wskazują na to, że aktorzy nie patrzą w kamerę, co może sugerować, że czytają tekst z kartki lub korzystają z promptera. „Nie patrzyła się w obiektyw, jakby czytała coś z kartki", „Wzrok ciągle w jednym kierunku, nerwowe przenoszenie ciężaru ciała oraz gest na koniec.", „Nie patrzyła w kamerę, czytała tekst wyświetlany obok".

Innym nienaturalnym elementem był sztuczny uśmiech osób występujących na obu nagraniach. „Wymuszony uśmiech, gestykulacja i ruchy ciała, czytanie tekstu, podnoszenie głosu na koniec niektórych fraz", „Sztuczny uśmiech na koniec, wypowiedz wyrecytowana jakby z pamięci". Nieudolna próba uśmiechania się mogła być spowodowana chęcią lepszego odbioru i bycia bardziej przekonującymi.

Ostatnim elementem, na który zwrócili swoją uwagę respondenci była niewiarygodność prezentowanych informacji. Niektóre odpowiedzi wskazują na nierealność prezentowanych treści, takich jak wysokość zarobków czy dostępność

usług, co może wpływać na postrzeganie autentyczności przekazu. „Nienaturalne wydaje się to, że facet ktory proponuje coś co przynosi mu „„„zyski”„ miałby chwalić sie tym oraz powodować wzrost konkurencyjnosci w swojej branży. Na ogół jest raczej odwrotnie. Do tego kwota 89tys wyswietlana na ekranie wydaje się nierealna przy wygladzie całego pomieszczenia jak i ubiorze aktora”, „pasek na filmie „„„zarobił 98 000 zł”„.

Podsumowując odpowiedzi respondentów dla filmów prezentujących nieznane osoby, w odpowiedzi na pytanie o nienaturalne elementy, obejmują głównie uwagi dotyczące sztuczności treści (ich wyuczenie na pamięć, czytanie z kartki, czytanie z promptera), brak autentyczności w zachowaniu i wypowiedzi oraz udawane gesty i uśmiechy. Ponadto kilkukrotnie zwrócono uwagę na nierealne obietnice zysków.

W ostatniej części niniejszego pod-podrozdziału analizie poddano odpowiedzi respondentów na pytanie dotyczące nienaturalnych elementów w prawdziwych filmach prezentujących średnio znanych influencerów (nagrania 4 i 5). Uwagi dotyczyły przede wszystkim nieudolnych prób bycia naturalnym ze swoim przekazem. „Najbardziej rzucalo się w oczy ucinanie filmiku i próba bycia naturalnym”, „Szybkie zdjęcie okularów zaraz na początku mówienia.”, „Wydawał sie trochę spięty – nie miał do końca flow na filmie”. Respondenci zauważają próby aktorów bycia naturalnymi, ale jednocześnie dostrzegają nienaturalne elementy, takie jak pretensjonalny styl bycia czy duża ilość przerw w mówieniu. Zwracana jest również uwaga na próby manipulacji tekstami lub nachalnego nakłaniania do inwestycji. „Może jestem uprzedzona do piramid finansowych i je wszędzie widzę, ale jego zachęcanie wręcz mnie zniechęca i widać, że będzie coś z tego miał, jak kogoś przekona.”, „Stwierdzenie, że niby jest świeżakiem. Jego wypowiedź wcale na to nie wskazuje.”, „manipulacyjne teksty”.

Pozostałe nienaturalne elementy, to przede wszystkim niespójności w wyglądzie. „Kryształowe zęby :)”, „Pretensjonalno-bucowaty styl bycia, machanie palcem w stronę rozmówcy”, „Wygląda jak Janusz biznesu”. Respondenci zwracają uwagę na nienaturalne elementy w wyglądzie aktorów, jak manipulacyjna stylizacja, co może wpływać na postrzeganie autentyczności prezentowanego materiału. Jeden z respondentów skomentował również otoczenie, w którym prowadzone było nagranie. „Przykro to stwierdzic, ale całość wyglada nienaturalnie (niemożliwie białe zęby, fryzura plus kto trzyma butelke whiskey na szafie?!)”.

Elementem, który nie pojawił się dotychczas we wcześniej omawianych odpowiedziach jest fakt, iż w odpowiedziach kilku respondentów wydaje się, że nie

zauważyli oni żadnych istotnych elementów nienaturalności w prezentowanych filmach, uznając je za naturalne. „Według mnie film został nagrany naturalnie", „Wszystko ok", „Nic nie było sztuczne, wydawał się bardzo naturalny", „Jest naturalnie".

Odpowiedzi respondentów sugerują, że nienaturalne elementy w wyświetlonych filmach średnio znanych influencerów obejmują próby bycia naturalnym, manipulacyjne teksty, nienaturalne elementy w wyglądzie i zachowaniu oraz brak autentyczności w wypowiedziach.

We wszystkich sześciu nagraniach powtarzającym się elementem były uwagi dotyczące niewizualnych aspektów filmów – głosu, jego tonu oraz nienaturalności. Elementem, który wyróżniał nieznane osoby na tle pozostałych była przede wszystkim sztuczność ich wypowiedzi oraz nieprofesjonalne nagranie. Obu tych aspektów udało się uniknąć przy filmach deepfake, przy których najmniej naturalnym elementem dla osób je oglądających zdają się być (oprócz głosu) wątpliwości co do ich wizerunku jako ekspertów inwestowania oraz pozostałych elementów z tym związanych (takich jak domniemany brak wiedzy czy przeczucie sponsorowanej reklamy). Co istotne, elementy te zdają się nie być dominujące w przypadku nagrań pozostałych, mniej znanych influencerów (filmy 4 i 5). W tych dwóch przypadkach najwięcej uwag dotyczyło szczegółów nienaturalnie wyglądającej twarzy oraz podejrzanej gry aktorskiej.

4.3.2 Rozpoznanie nagrań deepfake – uwagi dotyczące nienaturalnych elementów

Jak zaznaczono w powyższym podpodrozdziale, pierwsze pytanie – „jakie elementy aktora / aktorki były dla ciebie nienaturalne (sztuczne) na powyższym nagraniu?" miało na celu zebrać od respondentów ich subiektywne odczucia względem nienaturalnych elementów oraz informacje o tym, które składowe danego nagrania mogły świadczyć o jego podrobieniu. W niniejszej części analizy, zebrane zostały przykłady odpowiedzi na pierwsze pytanie otwarte osób, które twierdzą, iż rozpoznały nagrania deepfake spośród prezentowanych filmów. Jednak nie można określić czy osoba badana rozpoznała jedno, dwa nagrania, czy przez pomyłkę uznała za podejrzane również pozostałe filmy.

W pierwszej kolejności przeanalizowane zostały odpowiedzi respondentów na pytanie dotyczące nienaturalnych elementów w pierwszym filmie deepfake (film 1). Najwięcej uwag odnosi się do nienaturalność twarzy, co potwierdzać może, iż respondenci faktycznie rozpoznali deepfake na tym nagraniu. „mówienie znudzonym

głosem i twarz jak z generatora", „Twarz jest zmieniona", „Twarz aktorki", „Mimika twarzy", „Wygląda jakby miała nałożony jakiś filtr, który zmienia wygląd jej twarzy przez co nie można jednoznacznie stwierdzić czy jest to aktorka na której koncie ogląda się film.", „filtr, niezmienna mimika twarzy", „Chyba usta ma zrobione albo coś bo tak dziwnie rozmawia prawie nie ruszają sie jej usta a po za tym to dziwny filtr ma i jak rozmawia to troche ma styl taki depresyjny i to nie jest za fajne do słuchania". Wiele odpowiedzi koncentruje się na niewłaściwej mimice twarzy, sugerując, że wygląda ona sztucznie lub jakby była zmieniona przez filtr.

Ponadto licznie zauważona została sztuczność w wypowiedzi i intonacji. „Trochę sztucznie brzmi ta wypowiedź, bardziej też jak odczytana niż wypowiedziana naturalnie (z głowy).", „widoczne „„recytowanie",, wcześniej zapamiętanego tekstu (wygląda to nienaturalnie)", „mowa (intonacja), gestykulacja, uciekanie wzrokiem, dobor slow", „Nie naturalna mowa", „Cała wypowiedź wydawała się sztuczna", „Jej wypowiedź o walucie", „dziwny filtr ma i jak rozmawia to troche ma styl taki depresyjny i to nie jest za fajne do słuchania". Respondenci zwracają uwagę na sposób, w jaki aktorka wygłasza tekst, opisując go jako „recytowanie" lub „odczytane z głowy". Zauważają również brak naturalności w modulacji tonu głosu.

Jedna z badanych osób bezpośrednio kwestionuje prawdziwą tożsamość aktorki. „Wygląda jakby miała nałożony jakiś filtr, który zmienia wygląd jej twarzy przez co nie można jednoznacznie stwierdzić czy jest to aktorka, na której koncie ogląda się film". Tym samym respondent sugeruje, że film jest trudny do zidentyfikowania jako wykonany przez konkretnego aktora, ponieważ wygląda on inaczej niż zwykle.

Dwie odpowiedzi koncentrują się na nieprzekonującej mowie na temat inwestycji. „Zajmowanie się przez aktorkę/ modelkę, że złym PR tematem krypto który jest bardzo skomplikowanym zbiorem procesów.", „Często aktor/influencer reklamuje dany produkt/firmę tylko w celach zarobkowych sam nie wierząc w to". Poruszane jest również zagadnienie omawiania spraw inwestycyjnych na platformach społecznościowych. „omawianie na Instagramie/TiK Tok spraw biznesowych.", „nagrywa tik toki". Odpowiedzi te wskazują na niską autentyczność i merytoryczną wartość omawiania treści, takich jak inwestowanie czy sprawy biznesowe, przez influencerów na Instagramie lub TikToku.

Również w drugim z filmów deepfake (nagranie 6) zauważyć można liczne uwagi co do zastosowania filtrów twarzy. „twarz z użyciem filtrów", „Filtr na twarzy",

„Zmodyfikowana twarz.”, „Filtr + sztuczna wypowiedź”, „filtr, „„serdecznie polecam”„ było cóż. mało serdeczne i szczere”, „„Twarz aktora jest karykaturalna, plus filtr ze snapczata nie wzbudza zaufania, gdyż nie widać twarzy”„ „Filtry”. Tylko jeden respondent prawidłowo nazywa proces tworzenia nagrań: „filtr na twarzy (deepfake?)”. Ponadto niektórzy respondenci zwracają uwagę na fakt, że twarz aktora jest zakryta lub rozmazana, co może podważać jego wiarygodność. „Twarz aktora jest zakryta. Wygląda jak tania przeróbka z photoshopa”, „Twarz”, „Twarz aktora jest karykaturalna, plus filtr ze snapczata nie wzbudza zaufania, gdyż nie widać twarzy”.

Dla obu filmów deepfake odnaleźć można wspólne elementy, które mogły wpłynąć na rozpoznanie deepfake w tej grupie badanych. Jest to przede wszystkim rozpoznanie nałożonych filtrów, nienaturalnych zmian na twarzy oraz brak naturalnej mimiki twarzy czy inne zmiany wizualne. Ponadto w nieco mniejszym stopniu uwaga zwracana była na sposób wypowiedzi czy też jej treść. Jedynie dla kilku osób nienaturalne było to, iż o inwestowaniu wypowiada się znana osoba.

Dla porównania poniżej zamieszczono wypowiedzi respondentów dla filmów prezentujących nieznane osoby (nagrania 2 i 3). Podobnie jak w całym niniejszym podpodrozdziale, respondenci stwierdzili, iż rozpoznali nagrania deepfake. Nie jest jednak wiadome czy rozpoznali oba filmy oraz czy nie wzięli za nie któregoś z pozostałych nagrań.

Analizując odpowiedzi respondentów na pytanie dotyczące elementów nienaturalnych w prawdziwych filmach nieznanych osób, można zauważyć przede wszystkim zwracanie uwagi na czytanie tekstu przez osoby występujące. „Nienaturalna gestykulacja nieodpowiadająca komunikacji werbalnej, spoglądanie na tekst. Do tego nienaturalne akcentowanie.”, „pani czyta z kartki, która jest za kamerą i nienaturalnie się cieszy”, „sztuczny uśmiech, wzrok zwrócony w bok (prawdopodobne czytanie z kartki wcześniej przygotowanego tekstu)”, „Osoba nie mowiaca swoim zdanie tylko wyuczona na pamiec lub czytajaca z kartki”, „Wyraźnie widać, że czyta z kartki”, „czyta z jakiejś kartki obok obiektywu, sztucznie sie usmiecha”, „Aktor cały wypowiedz czyta z kartki”. Respondenci zauważają nienaturalność mimiki oraz w sposobie wypowiedzi, co sugerować ma, iż czytają oni tekst z kartki lub promptera znajdującego się obok kamery.

Ponadto niektórzy respondenci komentują, że zachowania aktorów wydają się wymuszone, przez co mimika twarzy i gesty są nienaturalne. „Wymuszona pogodnosc, uciekanie wzrokiem, zbyt szybka dykcja, ...”, „sztuczna mowa, tekst wyuczony

na pamiec, ograniczona mimika", „Powtórzenie przygotowanej przez kogoś wypowiedzi". Kilkukrotnie zwrócona zostaje uwaga na widoczne zdenerwowanie aktora. „Aktorka lekko zdenerwowana, zapewne pierwszy występ w sieci", „Aktor jest odrobinę zdenerwowany, widocznie nie ma wprawy w wypowiadaniu się przed kamerą". Stres oraz zdenerwowanie mają wpływać na ich sposób wypowiedzi i zachowanie przed kamerą.

Inni respondenci zwracają uwagę na kiepską prezentację wizualną osób występujących na omawianych nagraniach. „monotonny głos, pauzy, wiercenie sie na krzesle, marne tło i nieelegancki, nieformalny ubiór", „marne tło i nieelegancki, nieformalny ubiór", „Pokój dziecka". Zdaniem badanych osób ma to bezpośredni wpływ na brak zaufania u widza. „Byl naturalny normalnie ubrany, ale mowił nie wlasbymi slowami tylko nauczyl sie tekstu", „Zachowanie i prezentacja wizualna mówcy nie wzbudziła mojego zaufania".

Odpowiedzi respondentów, którzy twierdzą, iż rozpoznali deepfake dla filmów prezentujących osoby nieznane sugerują, że głównym nienaturalnym elementem było przede wszystkim czytanie tekstu, sztuczne gesty i mimika twarzy, zdenerwowanie aktora oraz nieprzekonująca wypowiedź. Zdaniem badanych osób może prowadzić to do postrzegania filmów jako nienaturalne lub sztuczne, a przez to nieprzekonywujące. Dodatkowo podkreślona została niedoskonałość prezentacji wizualnej – przede wszystkim ubioru osób oraz wygląd ich otoczenia.

Analizując odpowiedzi respondentów na pytanie dotyczące elementów nienaturalnych w prawdziwych filmach prezentujących średnio znanych influencerów (nagrania 4 i 5), zauważyć można kilka kluczowych obserwacji. Przede wszystkim jest to nienaturalność w ich gestach i zachowaniu. „nienaturalne gesty i słownictwo, zbyt na siłę, raczej to też słaba gra aktorska. Brakuje u aktora spójności między komunikacją werbalną i niewerbalną.", „mówi tak, jakby wcale nie wierzył w to co prezentuje", „Powtarzalność ruchów", „sztucznyt". Respondenci zwracają uwagę na nienaturalność gestów i zachowania influencerów, sugerując, że wydaje się być ono wymuszone lub zbyt na siłę.

Ponadto uwaga zwracana jest na brak autentyczności w wypowiedzi, zaznaczona przez wymuszone, stereoptypowe sformułowania. „cała ta gadka jest sztuczna", „mówi tak, jakby wcale nie wierzył w to co prezentuje", „Wrażenie jakby wszystko było mówione z pamięci z napisanego wcześniej tekstu", „zbyt dopingujace slowa, ale dykcja

i sposob przekazu ok". Wiele odpowiedzi wskazuje na to, że wypowiedzi influencerów są postrzegane jako sztuczne, a niekiedy nawet nieprzekonujące.

Inne z odpowiedzi zwracają uwagę na niewiarygodność filmów i marketingowe podejście twórców do zagadnienia. Niektórzy respondenci sugerują, że próby reklamowania produktów lub usług przez influencerów wydają się niewiarygodne, szczególnie gdy nie mają oni rzeczywistego związku z reklamowanymi produktami (w tym przypadku inwestycjami). „Chęć zareklamowania czegoś z czym powyższa osoba w rzeczywistości nie ma nic wspólnego.", „Tak nagle odzywa się do niego firma? jestem ciekaw, kiedy do mnie zadzwonią", „Osoba przedstawiony na nagraniu jest niezwiązana z tego typu rzeczami.", „Zajmowanie się tematem, którego się nie zna.", „Tak zwany lans". Kilkukrotnie osoby badane zwracają uwagę na brak kompetencji osób występujących na nagraniach w niniejszym temacie. „Zajmowanie się tematem, którego się nie zna", „Osoba nieogarnięta omawiająca krypto.", „Mowienie o sobie ze sie jest powarznym". Odpowiedzi te wskazują na brak wiedzy lub zrozumienia tematu przez aktorów, co może wpływać na odbiór ich autentyczności.

Tylko jedna osoba zwróciła uwagę na powtarzalność ruchów – „Powtarzalność ruchów", co sugerować może, że ich zachowanie jest skrypcie lub zostało wyreżyserowane. Inny respondent swoją uwagę skupił na zbyt szybkiej dykcji – „zbyt szybka dykcja", co prowadzić może do niezrozumienia tematu. Ponadto tylko jedna osoba komentuje profesjonalizm influencera jako osoby reklamującej produkty za wynagrodzeniem. „Wrażenie jakby wszystko było mówione z pamięci z napisanego wcześniej tekstu". Odpowiedź ta sugerować może, że tekst był wypowiadany zbyt dokładnie, co brzmiało dla respondenta nienaturalnie i mogło mieć wpływ na odbiór autentyczności nagrania.

Odpowiedzi respondentów na nagrania prezentujące średnio znanych influencerów sugerują, że zwracali oni uwagę przede wszystkim na takie elementy jak nienaturalność w gestach i zachowaniu, brak autentyczności wypowiedzi oraz niewiarygodność w reklamie. Mogą one prowadzić do postrzegania prawdziwych filmów jako nienaturalne, zainscenizowane lub sztuczne. Dodatkowo, szybka, perfekcyjna dykcja jak i kierowanie przekazu przez osobę nieznającą tematu również są podnoszone jako czynniki wpływające na odbiór autentyczności.

4.3.3 Podsumowanie analizy

Analizując odpowiedzi respondentów w podziale na grupy prezentowanych filmów – deepfake, nieznane osoby, średnio znani influencerzy – oraz z zastosowaniem zmiennej nominalnej E – „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?" dostrzec można kilka istotnych wniosków. Przede wszystkim osoby, które twierdzą, iż rozpoznały deepfake w dużo większej mierze komentowały nienaturalny wygląd twarzy – zastosowane rzekomo filtry, nienaturalne elementy oczu czy ust oraz inne artefakty obrazu. W przeciwieństwie do nich, grupa, która nie rozpoznała deepfake uwagę swoją skupiała na modulowaniu głosu, merytoryce wypowiedzi oraz innych metodach rzekomej manipulacji.

Odpowiedzi te można było znaleźć przede wszystkim pod oboma filmami deepfake, co sugerować może, iż respondenci faktycznie prawidłowo zidentyfikowali te nagrania jako fałszywe. Również w grupie, która nie rozpoznała deepfake, zwłaszcza pod filmem szóstym, liczne były opinie informujące o zastosowaniu różnego rodzaju filtrów. Osobny badane uznały je jednak najprawdopodobniej za naturalne i pomimo ich zauważenia nie wpłynęły one na percepcję przekazu.

Analizując odpowiedzi dla grup średnio znanych influencerów oraz osób nieznanych, podzielone względem zmiennej nominalnej E, zauważyć można, iż są one ze sobą spójne i odnoszą się do podobnych elementów. Sugerować to może, iż respondenci, którzy twierdzą, iż rozpoznali deepfake rzeczywiście rozpoznali go wyłącznie na nagraniach 1 i 6. Ponadto nienaturalność na tych nagraniach wynikała najczęściej z elementów takich jak mowa, gestykulacja czy treść przekazu. Weryfikacja tego w jakim zakresie wpływ na rozpoznanie deepfake miała znajomość danego influencera przeprowadzona została w dalszej części niniejszego rozdziału – 4.4 oraz 4.6.

W przypadku zestawienia ze sobą odpowiedzi, z grupy która nie rozpoznała deepfake spod filmów deepfake z pozostałymi (odrzucona zostaje wyłącznie grupa odpowiedzi osób które nie rozpoznały deepfake spod filmów deepfake), okazuje się, iż są one zbieżne z odpowiedziami udzielonymi pod filmami średnio znanych influencerów (nagrania 4 i 5). We wskazanych przypadkach respondenci odnoszą się przede wszystkim do sztuczności głosu, nadmiernej gestykulacji czy braku wiarygodności występujących osób w omawianym zakresie (inwestycji). Weryfikacja tego w jakim zakresie filmy deepfake były naturalne dla respondentów przeprowadzona została na początku niniejszego rozdziału, w podrozdziale 4.1.

4.4 Rozpoznanie osoby występującej na nagraniu

W badaniu uwzględniono pytanie odnoszące się bezpośrednio do rozpoznawania osoby na nim prezentowanej. Założono, iż ponieważ influencerzy występujący na prawdziwych nagraniach 4 i 5 są średnio rozpoznawalni (liczba osób obserwujących ich profile na platformie Instagram to około 500 tys. osób) poziom ich rozpoznania zbliżony będzie do szóstego z filmów (deepfake) który prezentował podobną liczbę obserwujących osób w tym okresie. Największa rozpoznawalność powinna cechować film pierwszy (również deepfake), który prezentował aktorkę.

Kwestią mniej oczywistą jest, jak oceniali znajomość respondenci, którzy nie rozpoznali oszustwa, czyli zastosowania deepfake. Założono, iż ich ocena w tym pytaniu dla nagrań 1 i 6 będzie wyższa od oceny przyznawanej przez osoby, które rozpoznały oszustwo.

Dla filmów 2 i 3 ocena respondentów powinna wynosić 1. Osoby prezentowane na tych nagraniach nie pełną bowiem roli publicznych, a ich aktywność w mediach społecznych jest znikoma i ogranicza się do interakcji społecznych z grupą najbliższych znajomych. Założono, iż mało prawdopodobne jest by jakikolwiek ankietowany znał lub kojarzył wyświetlaną osobę.

Celem sprawdzenia czy respondenci rozpoznają osoby prezentowane na nagraniach, postanowiono zadać następujące pytanie: „Czy znasz osobę wyświetlaną na nagraniu?”. Odpowiedzi były możliwe na 10 stopniowej skali, gdzie 1 to „w ogóle”, zaś 10 „bardzo”. Poniżej zaprezentowano statystyki opisowe do trzynastego pytania.

Statystyki opisowe

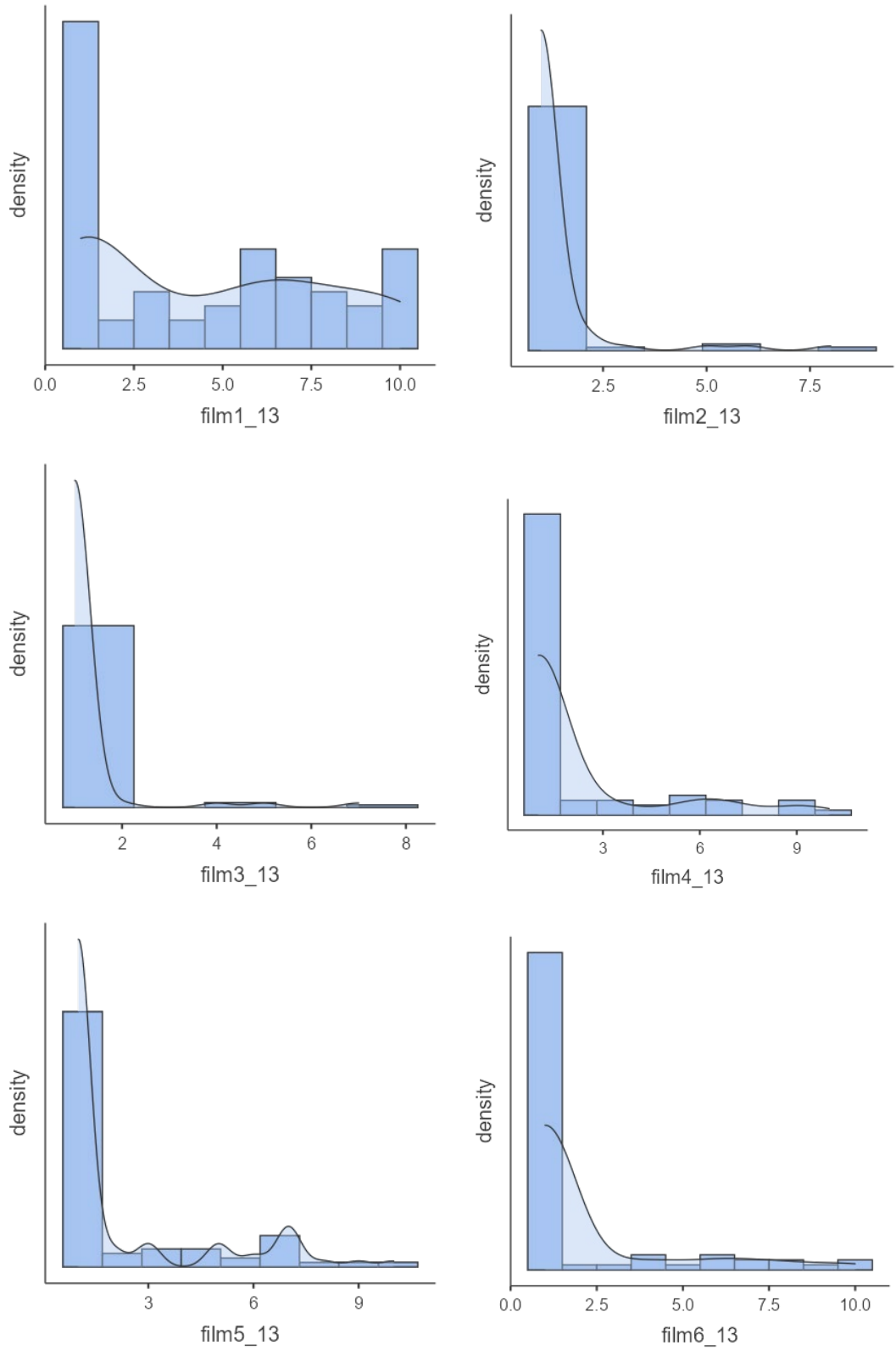
	film1_13	**film2_13**	**film3_13**	**film4_13**	**film5_13**	**film6_13**
N	60	80	79	80	80	79
Brakujące odpowiedzi	22	2	3	2	2	3
M	4.47	1.27	1.18	2.09	2.29	2.05
SE	0.436	0.121	0.0983	0.260	0.266	0.265
95% CI dolna granica przedziału ufności dla średniej	3.61	1.04	0.985	1.58	1.77	1.53
95% CI górna granica przedziału ufności dla średniej	5.32	1.51	1.37	2.60	2.81	2.57
Me	4.00	1.00	1.00	1.00	1.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	3.38	1.08	0.874	2.32	2.38	2.35
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	7.00	10.0	10.0	10.0

Statystyki opisowe

	film1_13	film2_13	film3_13	film4_13	film5_13	film6_13
SKE	0.340	4.82	5.44	2.11	1.68	2.19
SEk	0.309	0.269	0.271	0.269	0.269	0.271
K	-1.41	24.6	30.8	3.27	1.52	3.64
Std. error K	0.608	0.532	0.535	0.532	0.532	0.535
S-W	0.841	0.282	0.210	0.534	0.602	0.511
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 9 Statystyki opisowe dla trzynastego pytania „Czy znasz osobę wyświetlaną na nagraniu?”.

Dane z powyższej tabeli jednoznacznie pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$, wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności, a w większości przypadków, są one większe od 1 (z wyjątkiem filmu pierwszego) oraz wartości bezwzględne kurtozy, które w każdym przypadku są większe od wartości błędu standardowego tej miary. W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.



Wykres 10 Zbiór 6 wykresów odpowiedzi na trzynaste pytanie dla każdego z sześciu filmów.

Na wykresach 2, 3 i 6 widzimy przewagę niskich odpowiedzi. Uzyskane dane w tych przypadkach wskazują silną prawoskośność dla rozkładów, z znaczną dominacją

odpowiedzi 1 – wcale. Dzieje się tak zwłaszcza dla filmu 2 i 3, gdzie oba prezentowały wizerunki osób nieznanych. Nagranie 1 (deepfake) oraz 4 i 5 jest oceniane w podobny sposób. Część z respondentów zwraca uwagę, iż nagrania te mogą wpłynąć na odbiór wizerunku osób w nich występujących.

Aby zbadać, jak respondenci odpowiadali w zależności od rozpoznania deepfake, poniżej przedstawiono tabelę zawierającą dane analogiczne do poprzednich, z tym, że w tym przypadku rozkłady zmiennych zostały podzielone według zmiennej nominalnej E: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Statystyki opisowe

	zmienna nominalna E	**film1_13**	**film2_13**	**film3_13**	**film4_13**	**film5_13**	**film6_13**
N	nie	29	32	32	32	32	32
	tak	18	25	25	25	25	25
Brakujące odpowiedzi	nie	3	0	0	0	0	0
	tak	7	0	0	0	0	0
M	nie	5.14	1.19	1.22	2.38	2.53	2.91
	tak	3.72	1.48	1.16	1.80	2.48	1.32
SE	nie	0.650	0.130	0.189	0.423	0.440	0.522
	tak	0.749	0.337	0.160	0.458	0.542	0.229
95% CI dolna granica przedziału ufności dla średniej	nie	3.86	0.932	0.848	1.55	1.67	1.88
	tak	2.25	0.819	0.846	0.902	1.42	0.871
95% CI górna granica przedziału ufności dla średniej	nie	6.41	1.44	1.59	3.20	3.39	3.93
	tak	5.19	2.14	1.47	2.70	3.54	1.77
Me	nie	6.00	1.00	1.00	1.00	1.00	1.00
	tak	2.50	1.00	1.00	1.00	1.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	3.50	0.738	1.07	2.39	2.49	2.96
	tak	3.18	1.69	0.800	2.29	2.71	1.14
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00

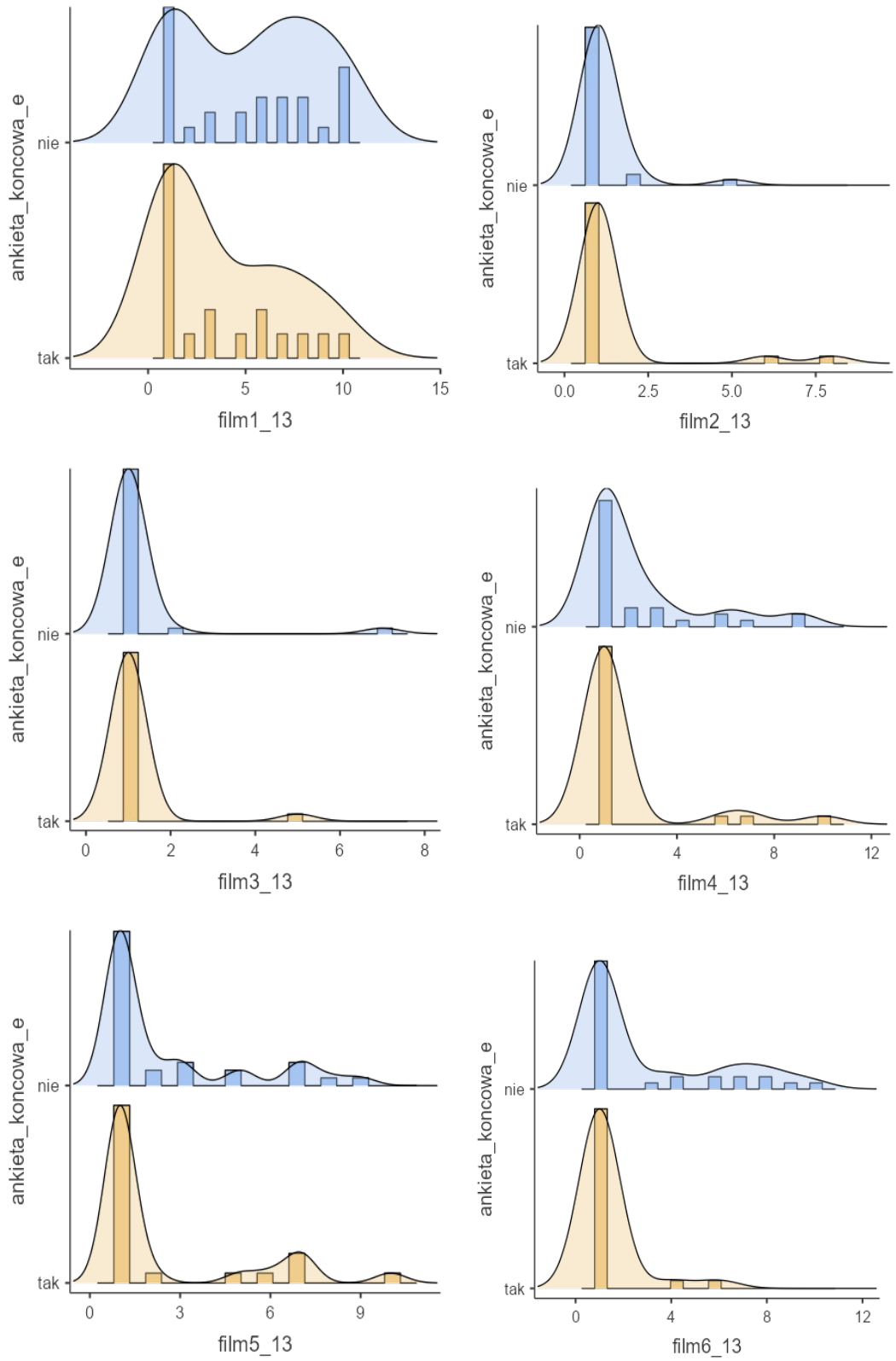
Statystyki opisowe

	zmienna nominalna E	film1_13	film2_13	film3_13	film4_13	film5_13	film6_13
Max	nie	10.0	5.00	7.00	9.00	9.00	10.0
	tak	10.0	8.00	5.00	10.0	10.0	6.00
SKE	nie	0.0320	4.82	5.44	1.83	1.47	1.20
	tak	0.756	3.47	5.00	2.84	1.60	3.65
SEk	nie	0.434	0.414	0.414	0.414	0.414	0.414
	tak	0.536	0.464	0.464	0.464	0.464	0.464
K	nie	-1.58	24.7	30.1	2.34	0.797	-0.0882
	tak	-0.911	11.3	25.0	7.41	1.28	13.2
Std. error K	nie	0.845	0.809	0.809	0.809	0.809	0.809
	tak	1.04	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.858	0.281	0.212	0.641	0.668	0.681
	tak	0.818	0.316	0.203	0.402	0.604	0.314
PS-W	nie	0.001	< .001	< .001	< .001	< .001	< .001
	tak	0.003	< .001	< .001	< .001	< .001	< .001

Tabela 10 Statystyki opisowe dla trzynastego pytania „Czy znasz osobę wyświetlaną na nagraniu?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości bezwzględne kurtozy, które w większości przypadków są większe od błędu standardowego tej miary, jak również wartości skośności, które w każdym przypadku (z wyłączeniem filmu pierwszego, odpowiedzi „nie") były wyższe od błędu standardowego skośności. Świadczą o tym również niskie wyniki (wszystkie filmy $p < 0,05$, a większość $p<0,001$) testu Shapiro-Wilk. W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Na poniższych wykresach przedstawiono histogramy dla każdego z filmów, z podziałem respondentów według zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").



Wykres 11 Zbiór 6 wykresów odpowiedzi na trzynaste pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na trzynaste pytanie, między osobami, które uważają, iż udało im się

rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zgodnie z założeniami, dla filmów 2 i 3, prezentujących nieznane osoby, obserwujemy znaczną dominację odpowiedzi 1 (nie znam). Również dla nagrań 4, 5 (średnio znani influencerzy) i 6 (deepfake) obserwujemy silną prawoskośność i zdecydowaną przewagę niskich odpowiedzi. W przypadku pierwszego filmu (deepfake) występuje zauważalna tendencja wyższych odpowiedzi w przypadku nierozpoznania deepfake niż w przypadku rozpoznania.

Celem sprawdzenia czy różnice te są istotne statystycznie przeprowadzono serię nieparametrycznych odpowiedników jednoczynnikowej analizy wariancji (ANOVA'y), czyli testów Kruskal-Wallis. Wyniki te zostały zamieszczone w poniższej tabeli. Zaobserwowano, iż jedynie w przypadku ostatniego filmu (deepfake) rozkłady różnią się między sobą w istotny statystycznie sposób ($p < 0,05$). Świadczy to o tym, że różnice pomiędzy tymi dwoma grupami są wystarczająco duże, aby móc je wiarygodnie wyjaśnić przy użyciu statystyk. Istotność statystyczna została użyta do potwierdzenia czy różnice między grupami są na tyle duże, aby nie były one przypadkowe i mogą być uważane za prawdziwe w sensie statystycznym.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_13	1.78936	1	0.181	0.03890
film2_13	0.00671	1	0.935	1.20e-4
film3_13	0.13988	1	0.708	0.00250
film4_13	3.53268	1	0.060	0.06308
film5_13	0.29911	1	0.584	0.00534
film6_13	5.85001	1	0.016	0.10446

Tabela 11 Test Kruskal-Wallis dla odpowiedzi do pytania trzynastego.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie trzynaste („Czy znasz osobę wyświetlaną na nagraniu?") będzie się różniła w przypadku różnych filmów, przeprowadzono test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 81,4; $p < 0,001$).

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na trzynaste pytanie, przeprowadzono test Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

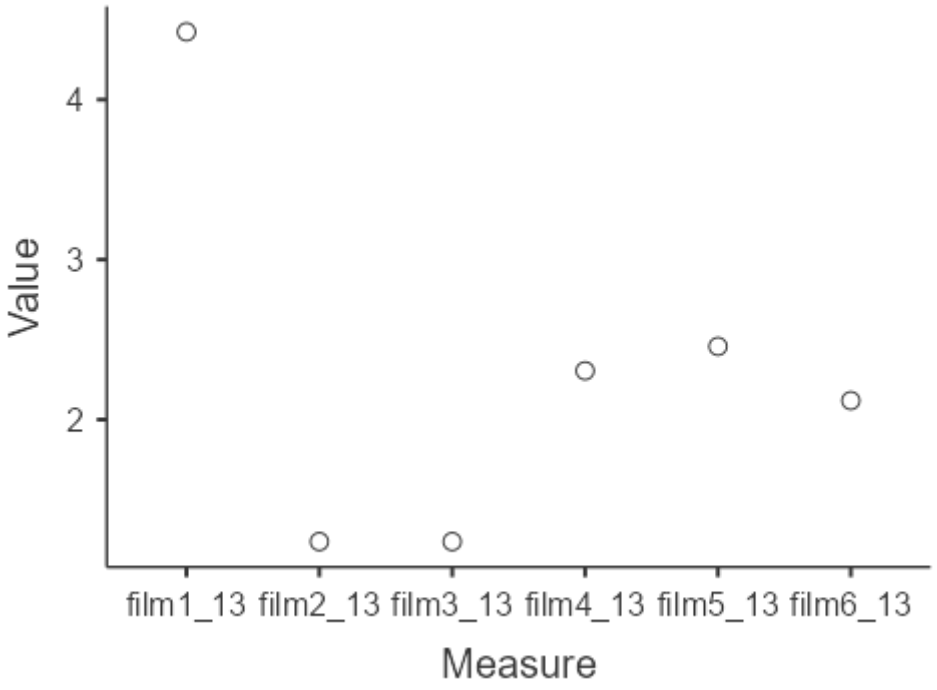
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_13	-	film2_13	8.835	< .001
film1_13	-	film3_13	9.309	< .001
film1_13	-	film4_13	5.861	< .001
film1_13	-	film5_13	5.473	< .001
film1_13	-	film6_13	6.206	< .001
film2_13	-	film3_13	0.474	0.636
film2_13	-	film4_13	2.974	0.003
film2_13	-	film5_13	3.361	< .001
film2_13	-	film6_13	2.629	0.009
film3_13	-	film4_13	3.448	< .001
film3_13	-	film5_13	3.835	< .001
film3_13	-	film6_13	3.103	0.002
film4_13	-	film5_13	0.388	0.698
film4_13	-	film6_13	0.345	0.731
film5_13	-	film6_13	0.733	0.464

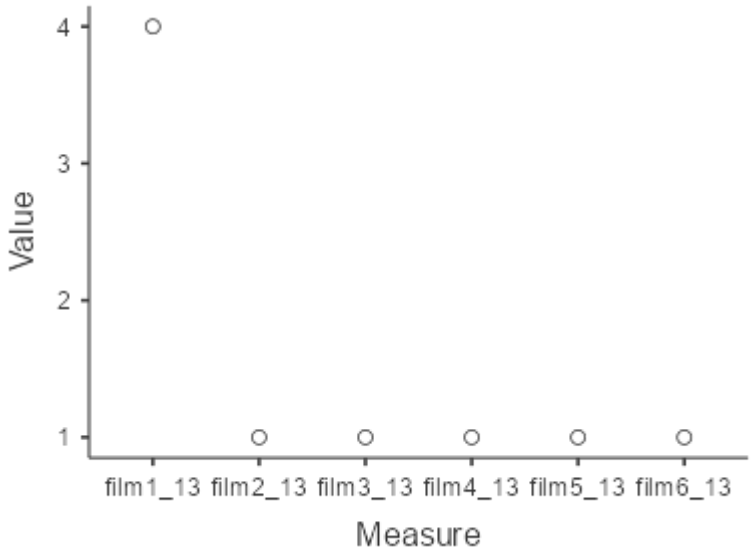
Tabela 12 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania trzynastego.

Powyższe obliczenia pokazały, że względem pytania trzynastego, film 1 różni się istotnie od każdego pozostałego filmu, czyli nagrań 2, 3, 4, 5 i 6, na co wskazuje $p < 0,001$. Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby, $p > 0,05$), natomiast różnica jest między nimi, a nagraniami 4, 5 oraz 6 (deepfake) $p < 0,05$. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) oraz 6 (deepfake) nie zaobserwowano różnicy ($p > 0,05$).

Wnioskować można, iż aktorka występująca w filmie 1 (deepfake) została rozpoznana istotnie różnie od osób występujących na pozostałych filmach. Istotnie różnie zostały również rozpoznane nieznane osoby, jak i średnio znani influencerzy (łącznie z średnio znanym influencerem z filmu szóstego). Celem weryfikacji tej hipotezy, poniżej umieszczono wykresy przedstawiające średnie oraz mediany. Zastosowanie mediany wynika z tego, iż rozkłady wyników są znacząco różne od rozkładu normalnego, przez co średnie są w tym przypadku mocno obciążone miarą tendencji centralnej.



Wykres 12 Średnia odpowiedzi dla pytania trzynastego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 13 Mediana odpowiedzi dla pytania trzynastego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Zgodnie z założeniami poczynionymi we wstępie do niniejszego podrozdziału, najbardziej znaną respondentom osobom okazała się aktorka z filmu pierwszego (deepfake). Co istotne, test Kruskal-Wallis dla zmiennej nominalnej E pytania pierwszego potwierdza, iż różnice pomiędzy grupą, która rozpoznała, a tą którą nie

rozpoznała deepfake nie są wystarczająco duże, aby móc mówić o różnicach w odpowiedziach między grupami. Może to świadczyć o dobrym przygotowaniu nagrań deepfake, gdyż lepsza znajomość aktorki mogła nie wpływać na rozpoznanie deepfake.

Mediana wyników dla filmów 2, 3, 4, 5 i 6 jednakowo przyjmuje wartość 1, natomiast ponieważ dzięki porównaniu Durbin-Conover wiadomo, iż odpowiedzi między grupami różnią się istotnie statystycznie, należy odnieść się do różnic w średnich. Filmy średnio znani influencerzy z filmów 4, 5 zostali istotnie statystycznie ocenieni jako bardziej znani, niż osoby nieznane. Podobnie stało się z filmem 6 (deepfake), który średnią oceny znajomości został zbieżnie rozpoznany jak influencerzy z filmów 4 i 5. Należy jednak zwrócić uwagę, iż dzięki testowi Kruskal-Wallis między grupami zmiennej nominalnej E, wnioskować można, iż na rozpoznanie deepfake mogła mieć znajomość lub nieznajomość tego influencera. Może to również świadczyć o gorszym przygotowaniu nagrania, od pierwszego filmu deepfake.

4.5 Zaufanie do wizerunku osoby występującej

Poniższy podrozdział porusza aspekty związane z zaufaniem do wizerunku prezentowanej na nagraniu osoby. Zastanawiające jest, czy wizerunek osób namawiających do fałszywych inwestycji wzbudzał zaufanie u osób je oglądających oraz które nagrania wzbudzały większe, a który mniejsze zaufanie. Czy osoby bardziej znane, występujące na nagraniach deepfake będą cieszyły się większym, czy mniejszym zaufaniem względem osób prezentowanych na pozostałych filmach? Czy osoby, które rozpoznały fałszywe filmy, będą je gorzej oceniać?

Pytanie, jakie zadano badanym brzmiało: „w jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?”. Odpowiedź udzielano na skali 1 do 10, gdzie 1 to w ogóle, 10 bardzo.

Zakładano, iż dla filmu pierwszego (deepfake) i filmów 4 oraz 5 (prawdziwe, influencerów) wynik zaufania do wizerunku będzie zbieżny. Natomiast w przypadku nagrania szóstego (deepfake) wynik oceny wizerunku będzie nieznacznie wyższy od wyniki z nagrań 2 i 3 (nieznane osoby). Wideo pierwsze było bowiem przygotowane z większą starannością niż film szósty. Wynikać to może z założenia, iż osoby, które nie rozpoznały deepfake nieco lepiej ocenią wizerunek prezentowanego influencera, natomiast te które rozpoznały manipulację wideo, nie powinny obdarzać zaufaniem nieznanego wizerunku. Poniżej zaprezentowano statystki opisowe dla wszystkich nagrań.

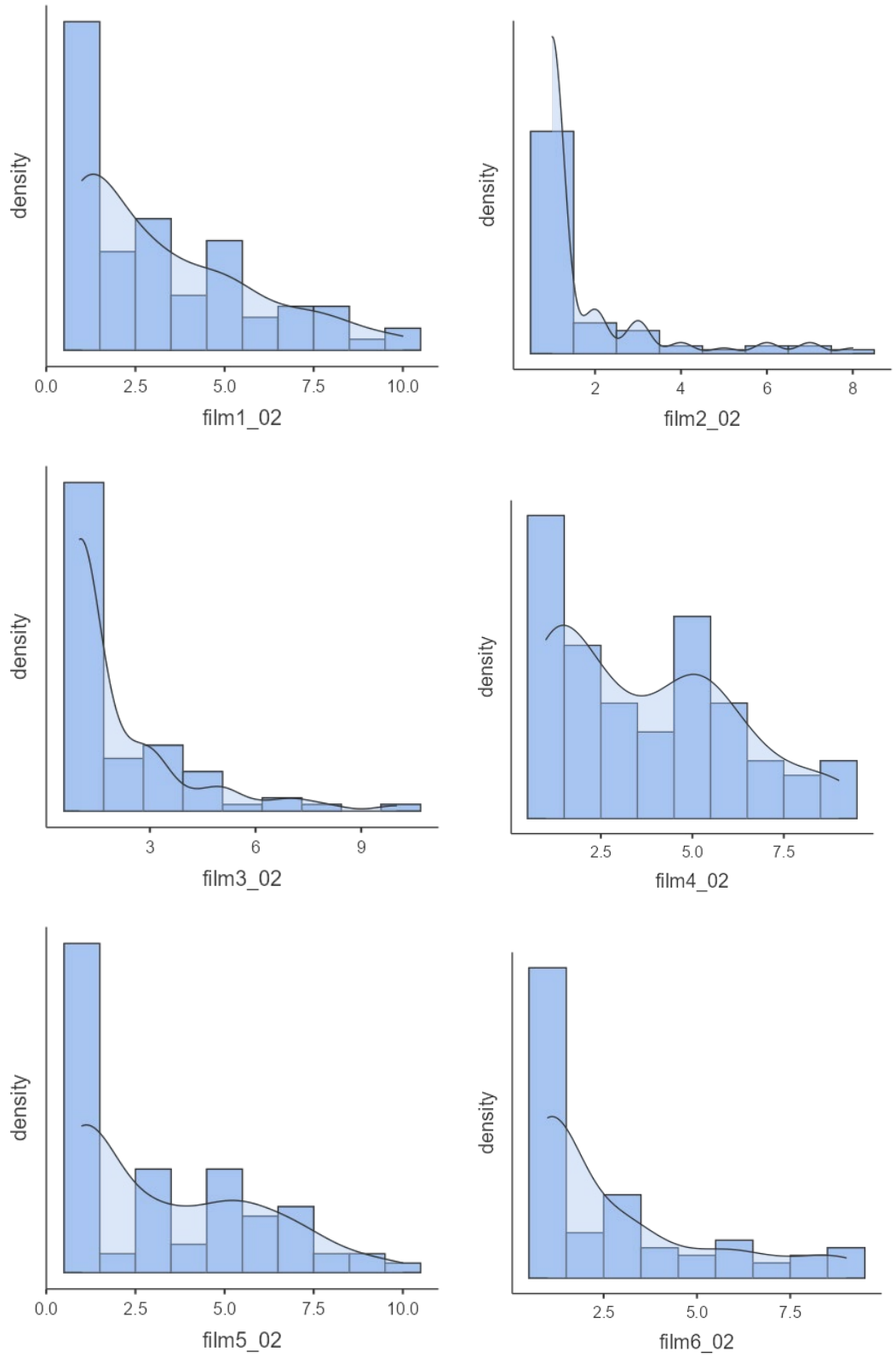
Statystyki opisowe

	film1_02	film2_02	film3_02	film4_02	film5_02	film6_02
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	3.26	1.74	2.05	3.74	3.35	2.80
SE	0.279	0.175	0.209	0.273	0.286	0.279
95% CI dolna granica przedziału ufności dla średniej	2.71	1.39	1.64	3.20	2.79	2.25
95% CI górna granica przedziału ufności dla średniej	3.81	2.08	2.46	4.27	3.91	3.34
Me	3.00	1.00	1.00	3.00	3.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	2.50	1.56	1.86	2.44	2.56	2.48
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	10.0	9.00	10.0	9.00
SKE	0.985	2.50	2.22	0.518	0.698	1.29
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	0.0572	5.83	5.09	-0.779	-0.679	0.474
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.842	0.544	0.637	0.897	0.837	0.745
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 13 Statystyki opisowe dla drugiego pytania „W jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?”.

Powyższa tabela przedstawia statystki opisowe dotyczące odpowiedzi na drugie (w kolejności ich prezentacji) pytanie: „w jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?”. Dane z powyższej tabeli jednoznacznie pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$ oraz wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności, a w trzech przypadkach, są one większe od 1. Świadczą o tym również wartości bezwzględne kurtozy, które we wszystkich przypadkach, z wyjątkiem nagrania szóstego, są większe od wartości błędu standardowego tej miary. W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Na poniższych wykresach znajdują się histogramy otrzymanych wyników – dla każdego filmu osobno – dla drugiego pytania („w jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?”). Histogramy również pokazują, iż rozkład wyników dla żadnego z filmów nie jest zbliżony do rozkładu normalnego.



Wykres 14 Zbiór 6 wykresów odpowiedzi na drugie pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi, zwłaszcza dla nagrań 2, 3 i 5. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Respondenci zdają się nie ufać wizerunkom osób występującym

na nagraniach, zarówno deepfake (pierwsze i ostatnie), jak i osób nieznanych (nagranie drugie i trzecie). Dopiero średnio znani influencerzy wzbudzili lekkie zaufanie, chociaż porównywalne z oceną filmu pierwszego. Wizerunek influencera z filmu czwartego został obdarzony największym zaufaniem. Drugi z influencerów, został obdarzony zaufaniem zbliżonym z zaufaniem do osoby występującej na pierwszym nagraniu (deepfake).

Celem weryfikacji czy rozpoznanie fałszu miało wpływ na ocenę zaufania względem osób na nagraniach, zdecydowano się na dalszą analizę z uwzględnieniem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Konieczne było bowiem sprawdzenie czy ocena wizerunku wśród respondentów, którzy rozpoznali fałsz, będzie statystycznie różna oraz jak wpływa to na ogólną ocenę wizerunku osób prezentowanych na nagraniach.

Tabela poniżej zawiera dane analogiczne do tych z poprzedniej analizy, ale w tym przypadku rozkłady zmiennych są podzielone według zmiennej nominalnej E, która brzmi: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Statystyki opisowe

	zmienna nominalna E	**film1_02**	**film2_02**	**film3_02**	**film4_02**	**film5_02**	**film6_02**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	4.59	1.63	1.72	3.94	3.63	3.94
	tak	2.36	1.96	2.20	3.36	3.08	1.64
SE	nie	0.521	0.209	0.221	0.442	0.439	0.539
	tak	0.360	0.422	0.451	0.476	0.490	0.282
95% CI dolna granica przedziału ufności dla średniej	nie	3.57	1.21	1.29	3.07	2.76	2.88
	tak	1.65	1.13	1.32	2.43	2.12	1.09

Statystyki opisowe

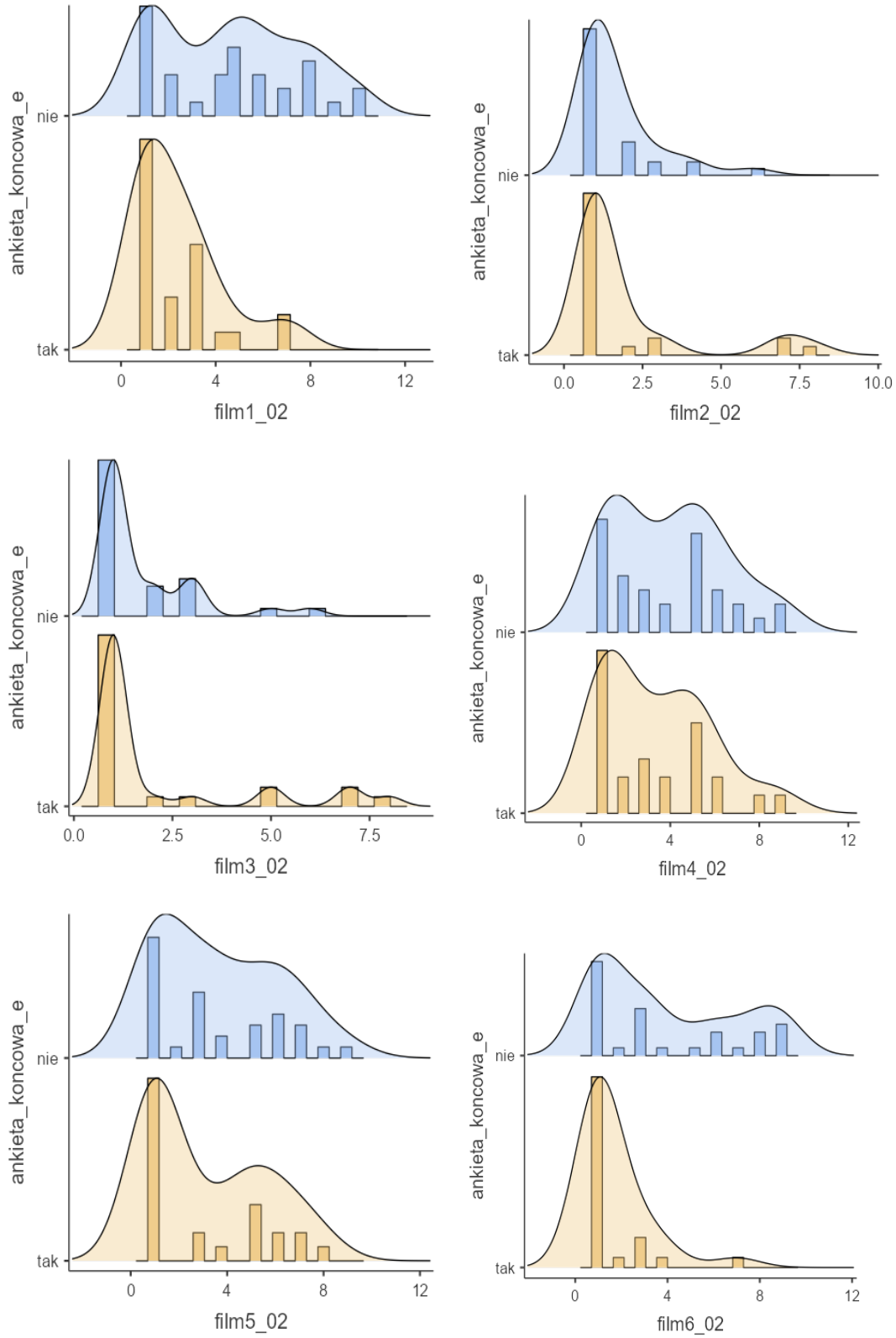
	zmienna nominalna E	film1_02	film2_02	film3_02	film4_02	film5_02	film6_02
95% CI górna granica przedziału ufności dla średniej	nie	5.62	2.04	2.15	4.80	4.49	4.99
	tak	3.07	2.79	3.08	4.29	4.04	2.19
Me	nie	5.00	1.00	1.00	4.00	3.00	3.00
	tak	2.00	1.00	1.00	3.00	1.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.95	1.18	1.25	2.50	2.49	3.05
	tak	1.80	2.11	2.25	2.38	2.45	1.41
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	6.00	6.00	9.00	9.00	9.00
	tak	7.00	8.00	8.00	9.00	8.00	7.00
SKE	nie	0.225	2.29	2.05	0.406	0.473	0.559
	tak	1.51	2.22	1.70	0.724	0.622	2.74
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.15	5.45	4.23	-0.832	-0.996	-1.27
	tak	1.81	3.70	1.46	-0.245	-1.18	8.34
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.911	0.606	0.641	0.908	0.881	0.825
	tak	0.756	0.517	0.594	0.871	0.790	0.532
PS-W	nie	0.012	< .001	< .001	0.010	0.002	< .001
	tak	< .001	< .001	< .001	0.004	< .001	< .001

Tabela 14 Statystyki opisowe dla drugiego pytania „W jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości skośności, które we wszystkich przypadkach, poza filmem 1 i 4 (grupa osób wskazująca „tak" względem zmiennej nominalnej E) są większe od błędu standardowego tej miary. Świadczą o tym również niskie wyniki dla wszystkich filmów ($p < 0,05$) testu Shapiro-Wilk (w większości przypadków $p < 0,001$). Ponadto wartości bezwzględne kurtozy są większe od błędu standardowego tej miary we

wszystkich filmach, z wyjątkiem filmu 4 (grupa osób wskazująca „tak” względem zmiennej nominalnej E).

Na poniższych wykresach przedstawiono histogramy dla każdego z filmów, z podziałem respondentów według zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”).



Wykres 15 Zbiór 6 wykresów odpowiedzi na drugie pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na drugie pytanie, między osobami, które uważają, iż udało im się

rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmów 1 i 6 (deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe. Najwyższy wynik zaufania do wizerunku zdają się mieć influencerzy występujący w nagraniach 4 i 5.

W celu sprawdzenia statystycznej istotności różnic przeprowadzono serię testów Kruskala-Wallisa. Wyniki znajdują się w tabeli poniżej. Statystycznie istotne różnice ($p < 0,05$) zaobserwowano tylko w przypadku pierwszego i ostatniego filmu (oba deepfake), co świadczy o tym, iż różnice te można wiarygodnie wyjaśnić statystycznie. Istotność statystyczna potwierdza, że różnice te nie są przypadkowe.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_02	8.37003	1	0.004	0.14946
film2_02	0.08680	1	0.768	0.00155
film3_02	0.00467	1	0.945	8.35e-5
film4_02	0.88990	1	0.346	0.01589
film5_02	0.88254	1	0.348	0.01576
film6_02	10.20692	1	0.001	0.18227

Tabela 15 Test Kruskal-Wallis dla odpowiedzi do pytania drugiego. Test czy osoby, które odpowiadały tak nie czy była między nimi różnica w ramach konkretnego filmu.

W powyższej tabeli widzimy, iż w przypadku filmu 1 (deepfake) istotność statystyczna testu Kruskal-Wallis jest mniejsza niż 0,05 w związku z czym powinniśmy odrzucić hipotezę zerową, która mówi o braku różnic między grupami, natomiast powinniśmy przyjąć hipotezę alternatywną mówiącą, iż te dwie grupy różnią się między sobą. W tym przypadku oznacza to, że osoby, które na pytanie E odpowiedziały „tak" w istotnie statystycznie różny sposób oceniały zaufanie względem wizerunku prezentowanego na nagraniu, w porównaniu, z osobami które odpowiedziały „nie". Podobnie odebrane zostało nagranie szóste (deepfake), którego p=0,001.

W przypadku filmu 2, 3, 4 i 5 istotność statystyczna testu Kruskal-Wallis jest większa niż 0,05. W związku z tym w przypadku tych filmów nie mamy podstaw do odrzucenia hipotezy zerowej i stwierdzenia, iż grupy te różnią się istotnie statystycznie.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie drugie („W jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?") będzie się różniła w przypadku różnych filmów, przeprowadzono test Friedmana. Test

ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 66,1; p < 0,001).

Celem zbadania różnic pomiędzy odpowiedziami na drugie pytanie w przypadku poszczególnych filmów, przeprowadzono serię testów Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_02	-	film2_02	5.280	< .001
film1_02	-	film3_02	4.315	< .001
film1_02	-	film4_02	1.769	0.078
film1_02	-	film5_02	0.482	0.630
film1_02	-	film6_02	1.823	0.069
film2_02	-	film3_02	0.965	0.335
film2_02	-	film4_02	7.049	< .001
film2_02	-	film5_02	5.763	< .001
film2_02	-	film6_02	3.458	< .001
film3_02	-	film4_02	6.085	< .001
film3_02	-	film5_02	4.798	< .001
film3_02	-	film6_02	2.493	0.013
film4_02	-	film5_02	1.287	0.199
film4_02	-	film6_02	3.592	< .001
film5_02	-	film6_02	2.305	0.022

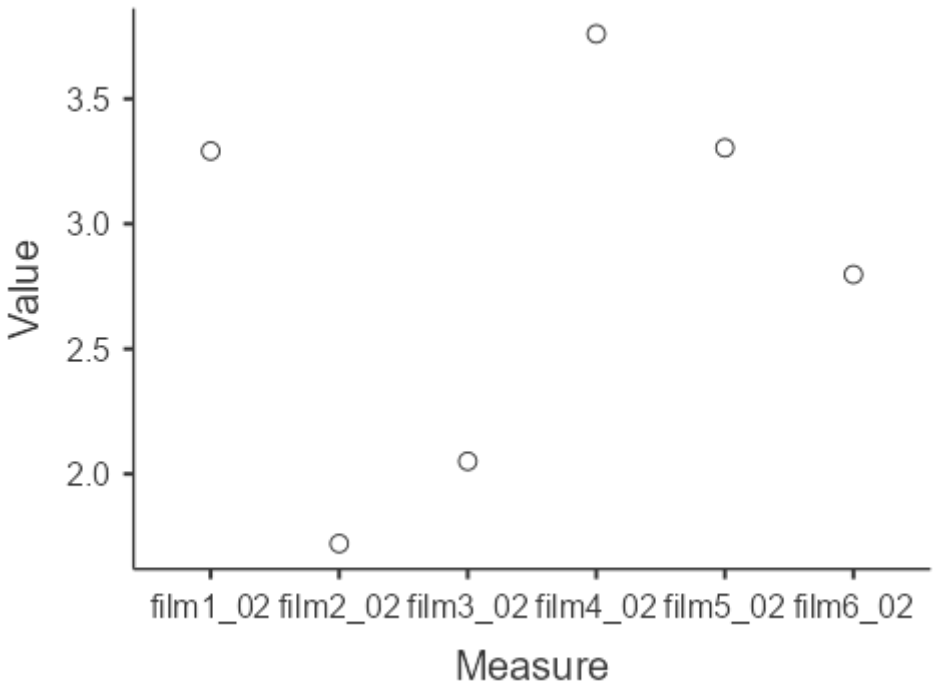
Tabela 16 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania drugiego. Osoby bardziej to czy tamto, nie biorąc pod uwagę odpowiedzi tak/nie.

Powyższe obliczenia pokazały, że film 1 różni się od filmu 2 i 3, na co wskazuje p < 0,05, natomiast nie zaobserwowano różnicy pomiędzy nim a filmem 4, 5 i 6 (p > 0,05). Potwierdza to wcześniejsze założenia, iż wizerunek influencera, prezentowany w filmie wykonanym z użyciem technologii deepfake, jest obdarzany zbliżonym zaufaniem przez osoby nieświadome oszustwa jak nagrania prawdziwych influencerów. Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4, 5 oraz 6. Respondenci większym zaufaniem obdarzają osoby znane od tych zupełnie nieznanych. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) nie zaobserwowano różnicy. Nie zaobserwowano jej również między filmami 5, a 6 (deepfake). Film 4 jest natomiast różny od filmu 6.

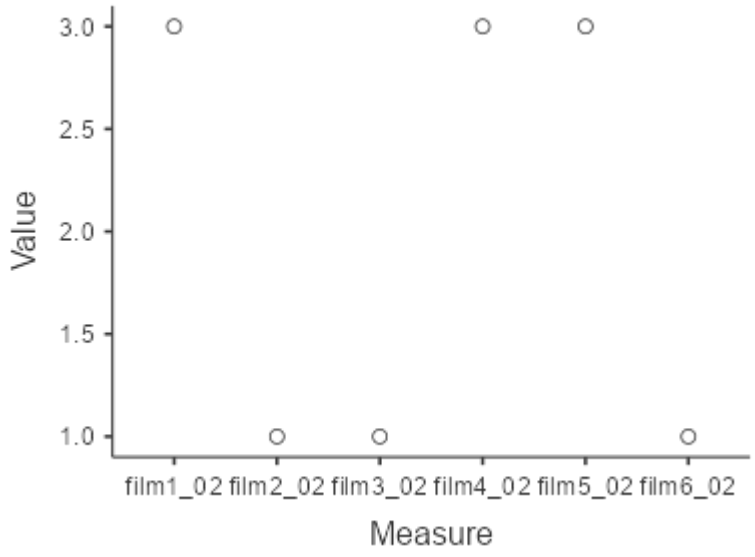
Wnioskować można, iż wizerunek osoby prezentowanej na filmie 1 (deepfake) nie był istotnie różny w odbiorze zaufania od influencerów z filmu 4 i 5, co świadczy

o ich podobnym odbiorze przez widzów. Również film 6 (deepfake) nie był istotnie różny od filmu 4.

W celu potwierdzenia tej hipotezy, poniżej zamieszczono wykresy z wartościami średnimi oraz medianami. Takie podejście jest konieczne, ponieważ rozkłady wyników są znacznie różne od rozkładu normalnego, co sprawia, że średnie mogą być obciążoną miarą tendencji centralnej. W związku z tym zaleca się zastosowanie mediany.



Wykres 16 Średnia odpowiedzi dla pytania drugiego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 17 Mediana odpowiedzi dla pytania trzeciego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Najwyżej oceniana średnia wyników zaufania do wizerunku osób występuje w przypadku filmu 4, 5 i 1. Niewiele niższy wynik ma nagranie szóste, jednak po analizie mediany, stwierdzić należy, iż podobnie jak wizerunki osób z nagrań 2 i 3, nie wzbudził on zaufania wśród respondentów.

Nagrania 1 (deepfake) oraz 4 i 5 są zauważalnie blisko oceniane, co widać zarówno po średniej jak i medianie (Me = 3). Ponadto z testu porównywania parami wiadomo, że filmy te nie były od siebie istotnie różne. Świadczyć to może o tym, iż duża część osób uwierzyła w prezentowany na nagraniu wizerunek. Ponadto, po rozkładzie przyznanych punktów stwierdzić należy, iż część osób która zadeklarowała, iż nie rozpoznała nagrań deepfake, znacząco wyżej oceniła zaufanie do wizerunku prezentowanej influencerki (deepfake) na pierwszym nagraniu, niż influencerów (osoby realne) z filmów 4 i 5.

4.6 Kojarzenie osoby występującej na nagraniu

Rozpoznawalność nagrań deepfake powinno rosnąc wprost proporcjonalnie do tego, kto w jakim stopniu zna osobę, pod którą ktoś inny próbuje się podszyć. Hipotetycznie dużo łatwiej jest rozpoznać nieścisłości w twarzy, mimiki czy głosu na nagraniu znanych nam osób niż tych których w ogóle nie kojarzymy. W niniejszym podrozdziale zbadane zostało, czy osoby z założenia rozpoznawalne rzeczywiście nimi są oraz czy osoby nieznane faktycznie nie zostały zidentyfikowane. Ponadto rozpoznanie

influencera, którego twarz została nałożona w nagraniach deepfake, spowodować może, iż respondenci rozpoznający w większym stopniu rozpoznają fałsz nagrania, niż osoby, które nie kojarzą danego influencera.

Celem weryfikacji tej hipotezy, badanym zadano następujące pytanie: „jak dobrze kojarzysz osobę prezentowaną na nagraniu?". Odpowiedzi udzielone zostały na dziesięciostopniowej skali, gdzie 1 oznaczało „w ogóle", zaś 10 „bardzo".

Założono, iż najwyższą ocenę zbierze pierwsze nagranie deepfake prezentujące twarz znanej influencerki. Poniżej jej wyniku, powinny się oceny osób występujących na nagraniach z filmów 4 i 5, ze względu na podobną rozpoznawalność osób w nich prezentowanych. Nieco gorzej rozpoznawanym powinien być influencer z drugiego filmu deepfake – nagrania 6. Charakteryzował się on zbliżoną do osób z filmów 4 i 5 liczbą obserwujących na portalu Instagram, jednak założono, iż niedoskonałości (artefakty) występujące na filmie po obróbce deepfake, obniżą jego rozpoznawalność. Nagrania 2 i 3 ze względu na prezentację nieznanych osób, powinny otrzymać najniższe wyniki, teoretycznie równe 1 („w ogóle"). Wizerunek osób na nich występujących nie był dotychczas nigdzie powszechnie prezentowany. W poniższej tabeli zaprezentowano statystyki opisowe z badania.

Statystyki opisowe

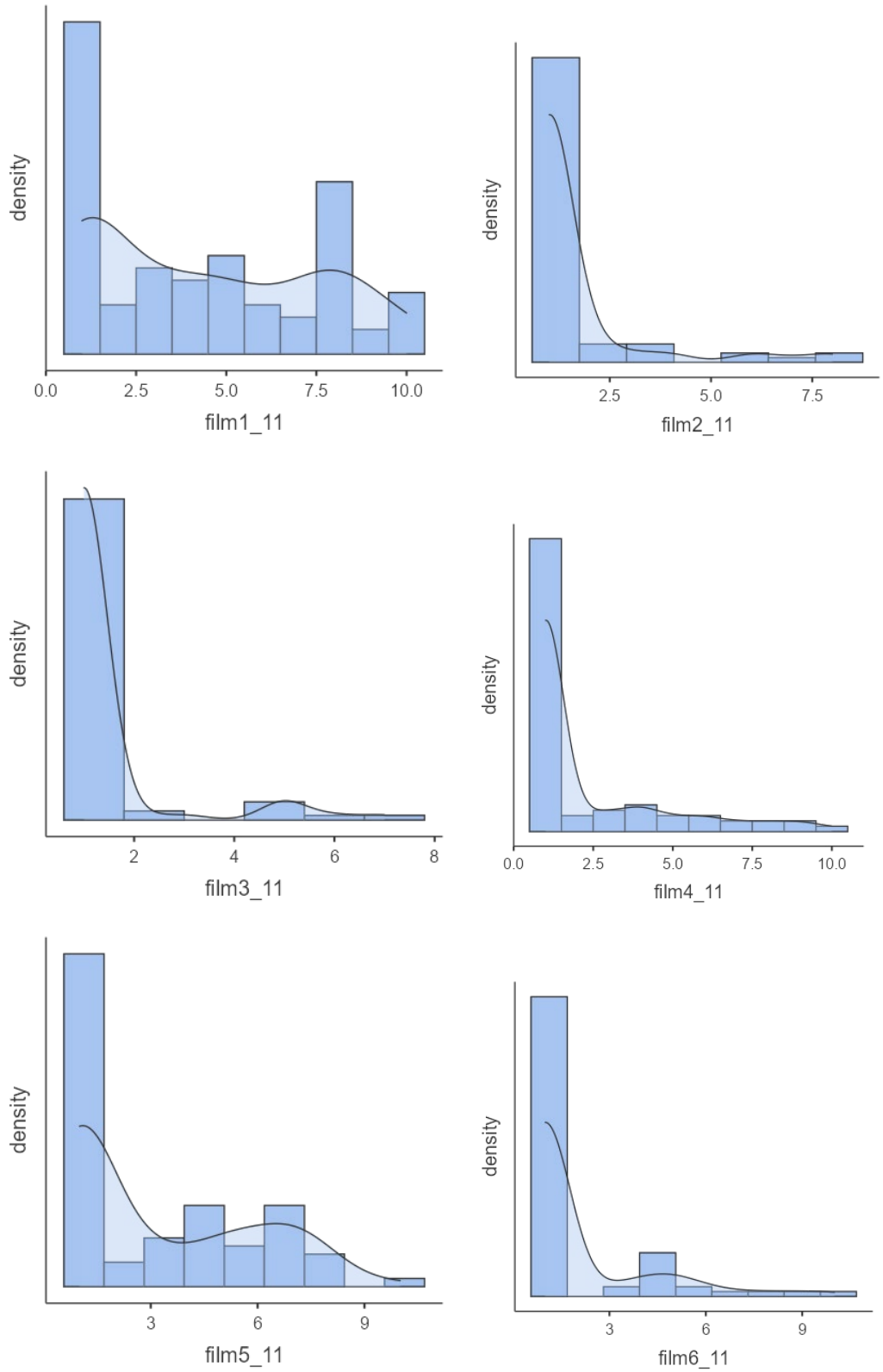
	film1_11	**film2_11**	**film3_11**	**film4_11**	**film5_11**	**film6_11**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	4.31	1.55	1.38	2.30	3.19	1.96
SE	0.348	0.173	0.139	0.261	0.296	0.232
`95% CI dolna granica przedziału ufności dla średniej	3.63	1.21	1.11	1.79	2.61	1.51
95% CI górna granica przedziału ufności dla średniej	4.99	1.89	1.65	2.81	3.77	2.42
Me	4.00	1.00	1.00	1.00	1.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	3.11	1.55	1.23	2.34	2.64	2.06
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	7.00	10.0	10.0	10.0
SKE	0.372	3.14	3.27	1.79	0.732	2.21
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-1.31	9.22	9.75	2.19	-0.949	4.33
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.862	0.408	0.343	0.627	0.780	0.538
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Statystyki opisowe

	film1_11	film2_11	film3_11	film4_11	film5_11	film6_11

Tabela 17 Statystyki opisowe dla jedenastego pytania „Jak dobrze kojarzysz osobę prezentowaną na nagraniu?”.

Dane przedstawione w powyższej tabeli jednoznacznie wskazują, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Potwierdzają to wyniki testu Shapiro-Wilk, który w każdym przypadku osiąga wartość $p < 0{,}001$, co jest najwyższym poziomem istotności statystycznej, powszechnie uznawanym w naukach społecznych. Dodatkowo, wartości skośności dla każdego filmu są wyższe niż błąd standardowy skośności, a w większości przypadków przekraczają 1 (jedynie skośności dla filmów 1 i 5 są poniżej 1). Ponadto, bezwzględne wartości kurtozy w każdym przypadku są większe niż błąd standardowy tej miary. W związku z brakiem normalności rozkładu uzyskanych wyników, w dalszych analizach statystycznych zastosowano testy nieparametryczne.



Wykres 18 Zbiór 6 wykresów odpowiedzi na jedenaste pytanie dla każdego z sześciu filmów

Na wykresach widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Dzieje się tak zwłaszcza dla filmu 2 i 3, gdzie obie prezentowane na nagraniach osoby były

osobami nieznanymi. Podobnie sytuacja prezentuje się z naganianiem 4, co jest poniekąd zaskakujące, gdyż prezentowało ono średnio znanego influencera z liczbą obserwujących na Instagramie ponad 400 tys. Niewiele więcej obserwujących miał średnio znany influencer z nagrania 5, którego profil obserwuje przeszło 600 tysięcy osób. Z nagrań deepfake najbardziej rozpoznawana wśród respondentów była influencerka z nagrania 1. Ponad połowa osób w jakimś stopniu rozpoznaje ją. Ostatnie z nagrań prezentujące średnio znanego influencera, pod względem rozpoznawalności, plasuje się pomiędzy nagraniami 4, a 5.

Poniższa tabela przedstawia dane analogiczne do poprzedniej, jednak w tym przypadku zmienne zostały podzielone według zmiennej nominalnej E, odnoszącej się do pytania: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania stworzone technologią deepfake?". Celem było sprawdzenie czy osoby, które rozpoznały fałsz, oceniały wyżej rozpoznawalność osoby z nagrania w porównaniu do tych, które nie dostrzegły manipulacji.

Statystyki opisowe

	zmienna nominalna E	**film1_11**	**film2_11**	**film3_11**	**film4_11**	**film5_11**	**film6_11**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	5.25	1.53	1.16	2.53	2.81	2.63
	tak	3.76	1.72	1.68	2.08	3.48	1.44
SE	nie	0.592	0.294	0.128	0.421	0.450	0.425
	tak	0.617	0.381	0.340	0.479	0.609	0.252
95% CI dolna granica przedziału ufności dla średniej	nie	4.09	0.954	0.906	1.71	1.93	1.79
	tak	2.55	0.973	1.01	1.14	2.29	0.946
95% CI górna granica przedziału ufności dla średniej	nie	6.41	2.11	1.41	3.36	3.69	3.46
	tak	4.97	2.47	2.35	3.02	4.67	1.93
Me	nie	6.00	1.00	1.00	1.00	1.00	1.00
	tak	3.00	1.00	1.00	1.00	1.00	1.00

Statystyki opisowe

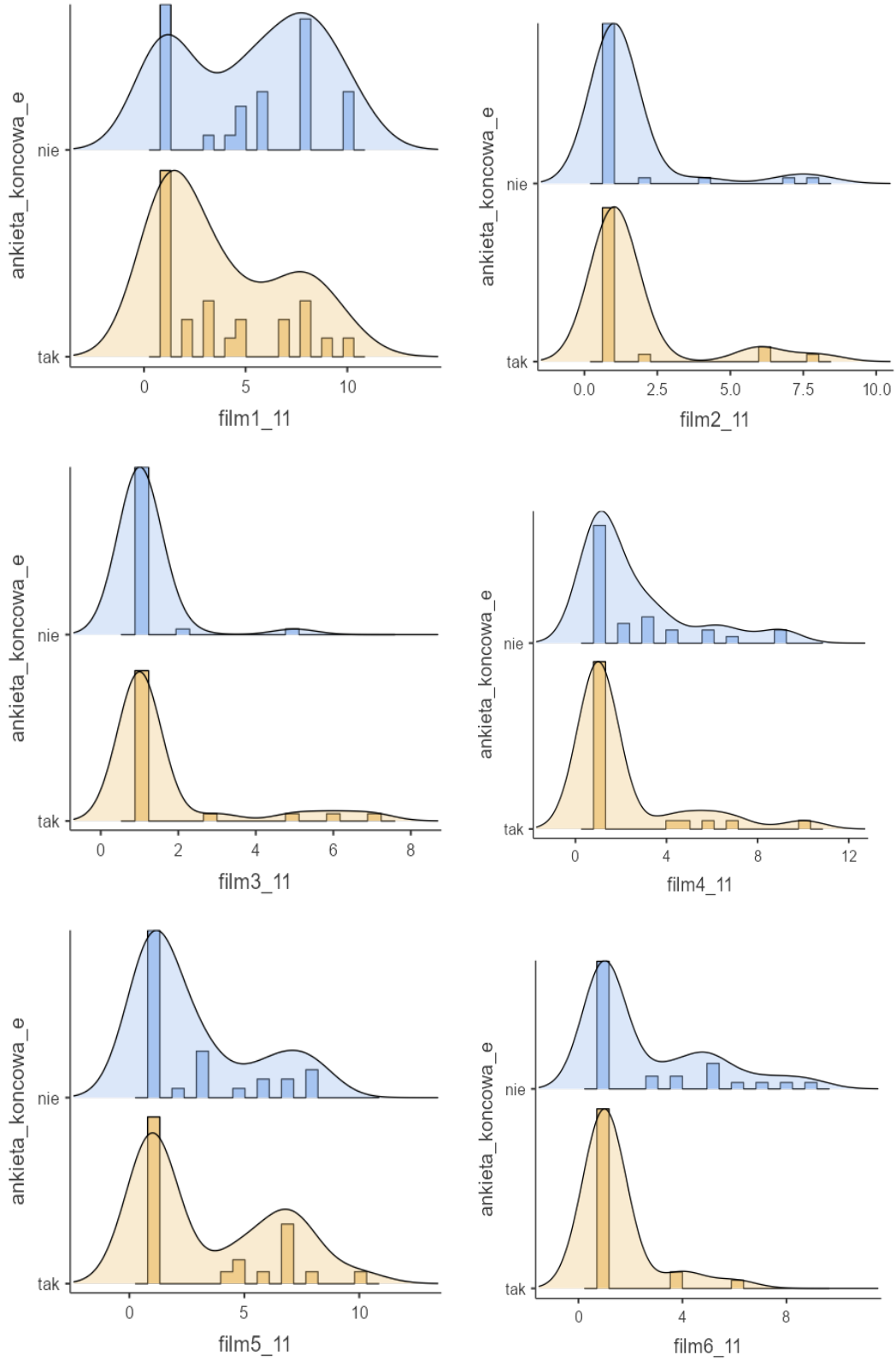
	zmienna nominalna E	film1_11	film2_11	film3_11	film4_11	film5_11	film6_11
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	3.35	1.67	0.723	2.38	2.55	2.41
	tak	3.09	1.90	1.70	2.40	3.04	1.26
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	8.00	5.00	9.00	8.00	9.00
	tak	10.0	8.00	7.00	10.0	10.0	6.00
SKE	nie	-0.150	3.32	5.21	1.67	1.11	1.26
	tak	0.716	2.60	2.42	2.25	0.638	2.86
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.50	10.4	28.0	1.91	-0.319	0.489
	tak	-1.02	5.67	4.76	4.49	-1.18	7.58
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.853	0.364	0.229	0.696	0.718	0.716
	tak	0.823	0.432	0.460	0.527	0.758	0.400
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 18 Statystyki opisowe dla jedenastego pytania „Jak dobrze kojarzysz osobę prezentowaną na nagraniu?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości kurtozy, gdzie w większości analizowanych przypadków wartość bezwzględna tej miary jest wyższa od jej błędu standardowego (dzieje się tak dla każdego filmu z wyjątkiem odpowiedzi „nie" dla filmów 5 i 6). Wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności, a w większości przypadków, są one większe od 1 (tylko wartość skośności filmu 1 oraz 5 były mniejsze od 1). Świadczą o tym również niskie wyniki ($p < 0{,}001$) testu Shapiro-Wilk wśród wszystkich nagrań.

Na poniższych histogramach zilustrowano wyniki dla każdego filmu, uwzględniając podział respondentów według zmiennej E („Jak uważasz, czy

w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”).



Wykres 19 Zbiór 6 wykresów odpowiedzi na jedenaste pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na jedenaste pytanie, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 6 (deepfake) zauważyć można delikatną przewagę wyższych odpowiedzi, dla grupy która nie rozpoznała, iż jest to nagranie fałszywe. W przypadku nagrania pierwszego nie zaobserwowano aż takich różnic.

Aby sprawdzić czy te różnice są istotne statystycznie przeprowadzono test Kruskal-Wallis. Wyniki te zostały zamieszczone w poniższej tabeli. Zgodnie z obserwacjami na powyższych wykresach, test Kruskal-Wallis potwierdza, iż jedynie w przypadku ostatniego filmu (deepfake) rozkłady różnią się między sobą w istotny statystycznie sposób ($p<0,05$). Świadczy to o tym, że różnice pomiędzy tymi dwoma grupami są wystarczająco duże, aby móc je wiarygodnie wyjaśnić przy użyciu statystyk.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_11	2.391	1	0.122	0.04270
film2_11	0.139	1	0.709	0.00248
film3_11	1.570	1	0.210	0.02803
film4_11	2.209	1	0.137	0.03945
film5_11	0.238	1	0.625	0.00426
film6_11	4.803	1	0.028	0.08577

Tabela 19 Test Kruskal-Wallis dla odpowiedzi do pytania jedenastego.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie jedenaste („Jak dobrze kojarzysz osobę prezentowaną na nagraniu?") będzie się różniła w przypadku różnych filmów, przeprowadzono test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 106; $p < 0,001$).

W celu analizy różnic pomiędzy odpowiedziami na pytanie jedenaste, zastosowano nieparametryczny odpowiednik testów post hoc. Wybrano test Durbin-Conover do przeprowadzenia porównań. Wyniki tego testu zostały przedstawione w poniższej tabeli.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_11	-	film2_11	8.82	< .001
film1_11	-	film3_11	9.95	< .001

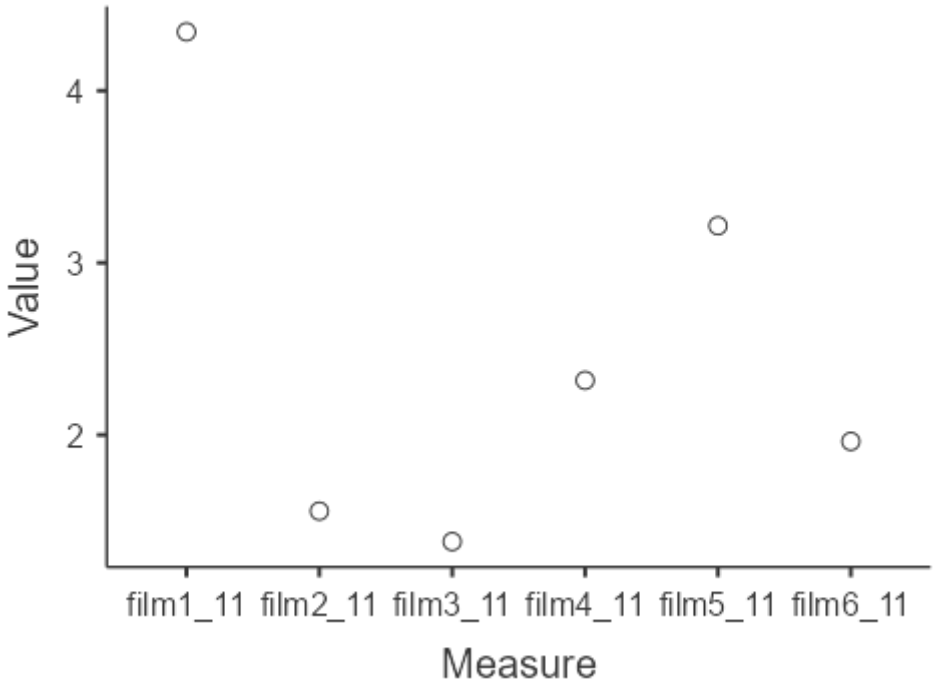
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_11	-	film4_11	6.44	< .001
film1_11	-	film5_11	2.94	0.003
film1_11	-	film6_11	7.51	< .001
film2_11	-	film3_11	1.14	0.257
film2_11	-	film4_11	2.37	0.018
film2_11	-	film5_11	5.88	< .001
film2_11	-	film6_11	1.30	0.194
film3_11	-	film4_11	3.51	< .001
film3_11	-	film5_11	7.01	< .001
film3_11	-	film6_11	2.44	0.015
film4_11	-	film5_11	3.51	< .001
film4_11	-	film6_11	1.07	0.286
film5_11	-	film6_11	4.57	< .001

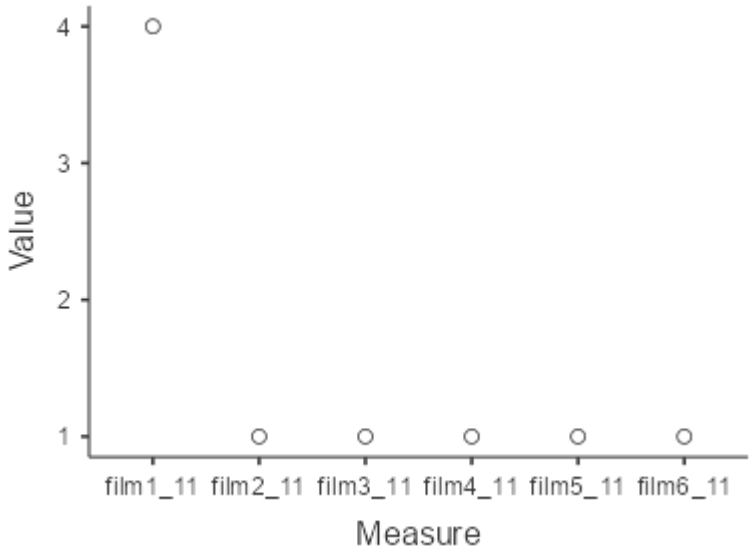
Tabela 20 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania jedenastego.

Powyższe obliczenia pokazały, że film 1 różni się istotnie statystycznie od pozostałych filmów. Natomiast nie zaobserwowano różnicy pomiędzy nagraniem 2, a filmem 3 i 6 ($p > 0,05$). Pomiędzy nagraniami 3, a 4 i 5 również zaobserwowane istotne statystycznie różnice. Nie zaobserwowano ich za to pomiędzy 3, a 6. Film 4 nie jest istotnie różny od filmu 5, za to jest od filmu 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 5, a 6 jednak nie zaobserwowano różnicy.

Wnioskować można, iż film 1 w największym stopniu różnił się od pozostałych filmów. Za to film 6 (deepfake) nie różnił się od filmów 2 i 3 co świadczy o ich podobnym rozpoznaniu przez widzów. Celem weryfikacji tej hipotezy, poniżej umieszczono również wykresy przedstawiające średnie oraz wykresy przedstawiające mediany. Z racji mocnego obciążenia średniej miarą tendencji centralnej, do analizy lepsze będzie zastosowanie mediany.



Wykres 20 Średnia odpowiedzi dla pytania jedenastego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 21 Mediana odpowiedzi dla pytania jedenastego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

We wstępie do poniższego podrozdziału przedstawiono założenie, iż ze względu na największą rozpoznawalność aktorki oraz najdokładniejsze przygotowanie filmu pierwszego (deepfake) osoba prezentowana na pierwszym nagraniu otrzyma najwyższe wyniki. Potwierdza to mediana, która wyraźnie pokazuje, iż osoba na nagraniu 1 była

w największym stopniu rozpoznawana (Me1 = 4). Co więcej, test Kruskal-Wallis dla tego filmu potwierdza, iż rozkłady nie różnią się między sobą w istotny statystycznie sposób ($p > 0,05$). Może to świadczyć o dobrym przygotowaniu nagrań deepfake, gdyż zarówno osoby rozpoznające influencerkę, jak i te nie kojarzące jej, w podobny sposób rozpoznawały deepfake.

Ponadto ze średniej wiemy, iż rozpoznanie influencera z filmu 6 (deepfake) oscylowało w obszarze rozpoznawania średnio znanego influencera z nagrania 4, a osoby nieznanej z filmu 2. W tym przypadku nagranie deepfake zostało nieco gorzej przygotowane. $P < 0,05$ testu Kruskal-Wallis pozwala wnioskować, iż znajomość wizerunku celebryty w istotny statystycznie sposób wpłynęła na możliwość rozpoznawania fałszu tego nagrania.

4.7 Zaufanie do filmowego przekazu aktorów

Chcąc ocenić wpływ filmów na jednostkę, nie należy patrzeć wyłącznie na ocenę wizerunku osoby na nim występującej. Zdarzyć się może, iż osoba z wyższym wstępnym poziomem zaufania, poprzez swoją mimikę, treść wypowiedzi, a także wykonywane gesty obniża je lub podwyższa. Technologia deepfake nie wpływa na podniesienie jakości tych elementów, natomiast rozpoznanie fałszu może zakłócić odbiór komunikatu lub go wzmocnić. Zastanawiające jest czy zmiana twarzy na danym nagraniu może wpłynąć na ocenę filmowego przekazu oraz jak oceniają filmowy przekaz osoby, które twierdzą, iż rozpoznały fałsz.

Poniższy podrozdział prezentuje wyniki analizy odpowiedzi na pytanie trzecie „Na ile jej (osoby występującej w filmie) filmowy przekaz wzbudza twoje zaufanie?”. Odpowiedź udzielano na skali 1 do 10, gdzie 1 to w ogóle, 10 bardzo.

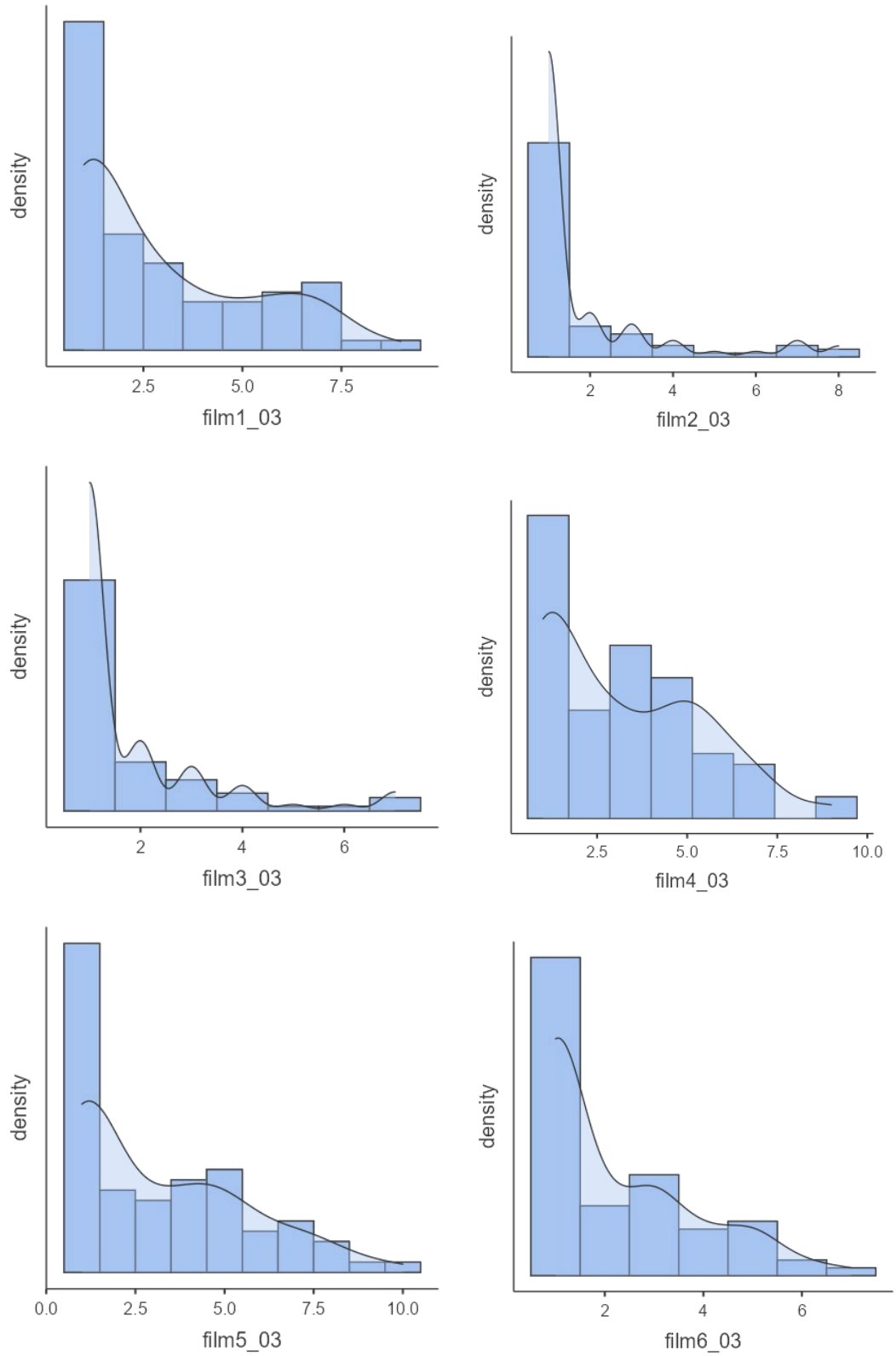
Podobnie jak w przypadku pytania drugiego („w jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?”) założono, iż film pierwszy (deepfake) otrzyma gorsze oceny niż filmy 4 oraz 5 (średnio znani influencerzy), ale będzie od nich różny. Założono również, iż ze względu na grę oraz mimikę osób występujących na nagraniach 2 i 3 zostaną one najgorzej ocenione. Nagranie szóste (deepfake) z racji ograniczonej gry aktorskiej powinno uplasować się w środkowym pułapie zaufania wraz z nagraniem pierwszym. Poniżej zaprezentowano statystki opisowe dla wszystkich nagrań.

Statystyki opisowe

	film1_03	film2_03	film3_03	film4_03	film5_03	film6_03
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	2.90	1.88	1.81	3.20	3.21	2.23
SE	0.251	0.197	0.167	0.246	0.269	0.177
95% CI dolna granica przedziału ufności dla średniej	2.41	1.49	1.48	2.72	2.69	1.88
95% CI górna granica przedziału ufności dla średniej	3.39	2.26	2.14	3.68	3.74	2.58
Me	2.00	1.00	1.00	3.00	2.50	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	2.25	1.76	1.49	2.20	2.41	1.58
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	9.00	8.00	7.00	9.00	10.0	7.00
SKE	0.951	2.29	2.21	0.681	0.834	1.10
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-0.364	4.53	4.60	-0.506	-0.273	0.228
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.808	0.567	0.611	0.868	0.848	0.777
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 21 Statystyki opisowe dla trzeciego pytania „Na ile jej filmowy przekaz wzbudza twoje zaufanie?”.

Powyższa tabela prezentuje statystyki opisowe dotyczące odpowiedzi na trzecie pytanie: „Na ile jej filmowy przekaz wzbudza twoje zaufanie?”. Analiza danych z tabeli jednoznacznie wskazuje, że nie można założyć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Dowodzą tego wartości testu Shapiro-Wilk, które dla każdego przypadku wynoszą $p < 0,001$, a także wartości skośności, które dla wszystkich filmów są wyższe od błędu standardowego skośności, przy czym w połowie przypadków przekraczają 1. Z uwagi na brak normalności rozkładu wyników, w kolejnych analizach statystycznych zastosowano testy nieparametryczne. Na poniższych wykresach przedstawione są histogramy (rozkłady) uzyskanych wyników dla każdego filmu osobno, dotyczące pierwszego pytania: „Na ile wyświetlony film wyglądał dla Ciebie naturalnie?”. Histogramy te również potwierdzają, że rozkład wyników dla żadnego z filmów nie jest zbliżony do rozkładu normalnego.



Wykres 22 Zbiór 6 wykresów odpowiedzi na trzecie pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi, zwłaszcza dla nagrań 2 i 3. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Respondenci zdają się nie ufać w filmowy przekaz osób

występujących na nagraniach, zarówno deepfake (film pierwszy i ostatni), jak i osób nieznanych (nagranie drugie i trzecie). Dopiero filmowy przekaz średnio znanych influencerów wzbudził wyższe zaufanie. Filmowy przekaz prezentowany na nagraniu 4 został obdarzony największym zaufaniem. Drugi z średnio znanych influencerów – na nagraniu 5 – został obdarzony zaufaniem zbliżonym z zaufaniem do osoby występującej na pierwszym nagraniu.

Celem weryfikacji czy odpowiedzi na pytanie dotyczące zaufania do filmowego przekazu, były różne w zależności od tego czy respondent rozpoznał czy nie rozpoznał nagrania deepfake, zdecydowano się na dalszą analizę z uwzględnieniem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Jest to bowiem niezbędne do ustalenia czy zmiana twarzy na danym nagraniu wpływa na ocenę filmowego przekazu zwłaszcza przez osoby, które twierdzą, iż rozpoznały fałsz.

Poniższa tabela przedstawia analogiczne dane jak prezentowana powyżej, natomiast w tym przypadku rozkłady zmiennych zostały podzielone względem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Statystyki opisowe

	zmienna nominalna E	**film1_03**	**film2_03**	**film3_03**	**film4_03**	**film5_03**	**film6_03**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	4.00	1.56	1.50	3.25	3.38	2.88
	tak	1.92	1.92	1.88	3.12	3.28	1.48
SE	nie	0.460	0.229	0.156	0.391	0.396	0.317
	tak	0.321	0.428	0.362	0.477	0.552	0.209
95% CI dolna granica przedziału ufności dla średniej	nie	3.10	1.11	1.20	2.48	2.60	2.25
	tak	1.29	1.08	1.17	2.18	2.20	1.07

Statystyki opisowe

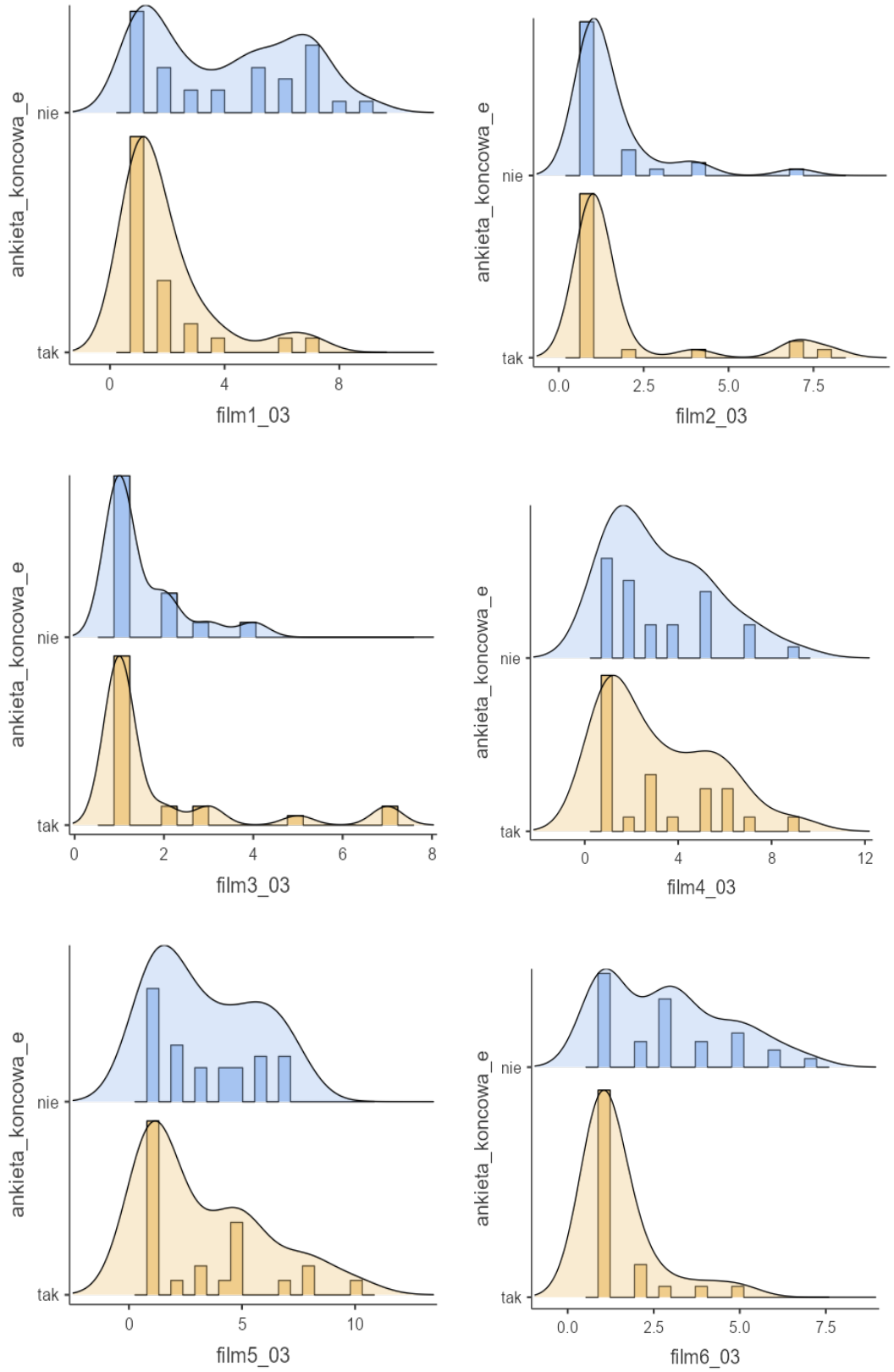
	zmienna nominalna E	film1_03	film2_03	film3_03	film4_03	film5_03	film6_03
95% CI górna granica przedziału ufności dla średniej	nie	4.90	2.01	1.80	4.02	4.15	3.50
	tak	2.55	2.76	2.59	4.06	4.36	1.89
Me	nie	4.00	1.00	1.00	2.50	3.00	3.00
	tak	1.00	1.00	1.00	3.00	2.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.60	1.29	0.880	2.21	2.24	1.79
	tak	1.61	2.14	1.81	2.39	2.76	1.05
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	9.00	7.00	4.00	9.00	7.00	7.00
	tak	7.00	8.00	7.00	9.00	10.0	5.00
SKE	nie	0.199	3.00	1.82	0.841	0.396	0.595
	tak	2.18	2.22	2.21	0.826	0.961	2.43
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.42	9.99	2.57	-0.0706	-1.39	-0.637
	tak	4.45	3.56	3.98	-0.294	-0.107	5.56
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.882	0.511	0.627	0.874	0.857	0.879
	tak	0.641	0.487	0.556	0.835	0.805	0.539
PS-W	nie	0.002	< .001	< .001	0.001	< .001	0.002
	tak	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 22 Statystyki opisowe dla trzeciego pytania „Na ile jej filmowy przekaz wzbudza twoje zaufanie?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to między innymi wartości bezwzględne kurtozy, które z wyjątkiem filmu czwartego oraz grup odpowiedzi względem zmiennej nominalnej E „tak" w nagraniu piątym oraz „nie" w filmie szóstym, przyjmują wartości wyższe od błędu standardowego tej miary. Wskazują na to również wartości skośności, które z wyjątkiem grupy „nie" w pytaniu pierwszym oraz „tak" w pytaniu piątym, są większe

od błędu standardowego skośności. O braku rozkładu normalnego świadczą również niskie wyniki ($p < 0,05$) testu Shapiro-Wilk (w większości przypadków $p < 0,001$).

Na poniższych wykresach przedstawiono histogramy dla każdego filmu, z uwzględnioną zmienną nominalną E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”).



Wykres 23 Zbiór 6 wykresów odpowiedzi na trzecie pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na pytanie trzecie, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 1 oraz 6 (oba deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe. Świadczyć to może o tym, iż rozpoznanie fałszu wpłynęło na ogólną ocenę nagrania. Osoby te w mniejszym stopniu dały się przekonać filmowemu przekazowi niż osoby, które fałszu nie rozpoznały. Wśród tych rozkład odpowiedzi zdaje się wyglądać zbieżnie z rozkładem odpowiedzi wśród filmów 4 i 5. Najsłabiej ocenione zostało zaufanie do filmowego przekazu dwóch, nieznanych osób z nagrań 2 i 3.

Aby zbadać istotność statystyczną różnic, zastosowano test Kruskala-Wallisa. Wyniki zamieszczono w poniższej tabeli. Tylko w przypadku pierwszego i ostatniego filmu (oba deepfake) zaobserwowano statystycznie istotne różnice ($p < 0,05$), co oznacza, że różnice te są na tyle duże, że można je uznać za wiarygodne.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_03	9.53118	1	0.002	0.17020
film2_03	0.03885	1	0.844	6.94e-4
film3_03	0.00159	1	0.968	2.84e-5
film4_03	0.20638	1	0.650	0.00369
film5_03	0.36780	1	0.544	0.00657
film6_03	10.94800	1	< .001	0.19550

Tabela 23 Test Kruskal-Wallis dla odpowiedzi do pytania trzeciego.

Z powyższej tabeli wnioskować można, iż respondenci deklarujący rozpoznanie nagrań deepfake, odpowiadali na to pytanie w odmienny sposób niż osoby, które fałszu nie rozpoznały. Na podstawie powyższych wykresów stwierdzić można, iż osoby te oceniały zaufanie do filmowego przekazu, na rozpoznanych przez siebie nagraniach znacznie gorzej niż to prezentowane na nagraniach 4 i 5, jednak lepiej niż na filmach 2 i 3.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie trzecie („na ile jej filmowy przekaz wzbudza twoje zaufanie?") będzie się różniła w przypadku różnych filmów, przeprowadzono test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 55,5; $p < 0,001$).

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na pytanie trzecie, przeprowadzono test Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

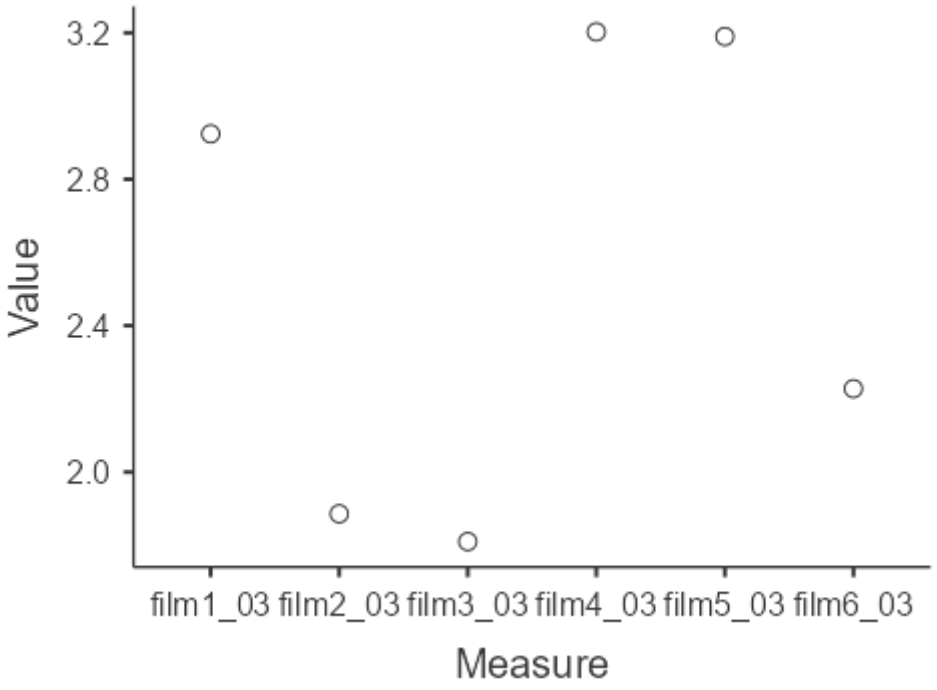
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_03	-	film2_03	3.834	< .001
film1_03	-	film3_03	4.410	< .001
film1_03	-	film4_03	1.397	0.163
film1_03	-	film5_03	1.150	0.251
film1_03	-	film6_03	2.191	0.029
film2_03	-	film3_03	0.575	0.566
film2_03	-	film4_03	5.231	< .001
film2_03	-	film5_03	4.985	< .001
film2_03	-	film6_03	1.643	0.101
film3_03	-	film4_03	5.806	< .001
film3_03	-	film5_03	5.560	< .001
film3_03	-	film6_03	2.218	0.027
film4_03	-	film5_03	0.246	0.805
film4_03	-	film6_03	3.588	< .001
film5_03	-	film6_03	3.341	< .001

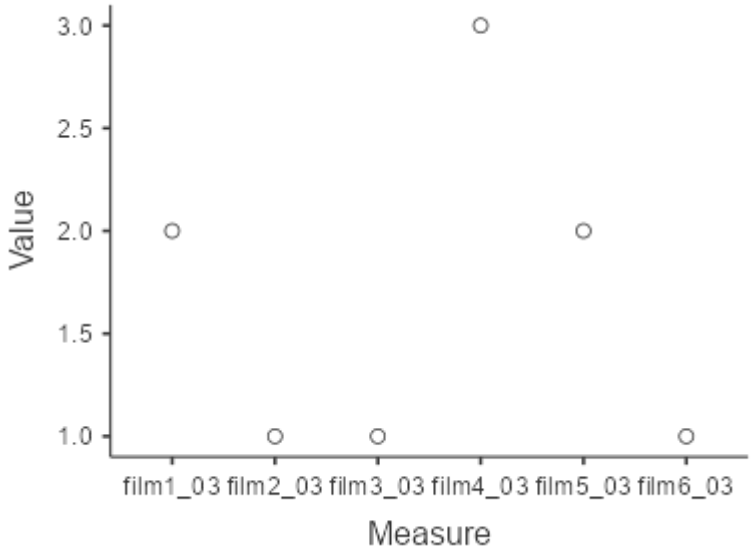
Tabela 24 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania trzeciego.

Powyższe obliczenia pokazały, że film 1 różni się od filmu 2, 3 i 6, na co wskazuje $p < 0,05$, natomiast nie zaobserwowano różnicy pomiędzy nim a filmem 4 i 5 ($p > 0,05$). Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4 oraz 5. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również nie zaobserwowano różnicy. Natomiast film 6 różni się zarówno od nagrań 4 i 5 jak i 3. Natomiast nie zaobserwowano różnic między nim, a filmem 2.

Na podstawie porównania parami wnioskować można, iż film 1 nie był istotnie różny od filmów 4 i 5, co świadczy tym, iż filmowy przekaz nagrania deepfake może wzbudzać podobne zaufanie, jak w nagraniach influencerów. Aby zweryfikować tę hipotezę, poniżej zaprezentowano zarówno wykres średnich, jak i wykres median. Różnice te wynikają z istotnego odchylenia rozkładów wyników od rozkładu normalnego. W związku z tym średnie są obciążoną miarą tendencji centralnej, co uzasadnia stosowanie mediany jako bardziej adekwatnej miary.



Wykres 24 Średnia odpowiedzi dla pytania trzeciego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 25 Mediana odpowiedzi dla pytania trzeciego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Średnia pokazuje, iż najwyżej punktowane były filmy 4 i 5, a za nimi niewiele niżej film 1 (deepfake). Po medianie widać, że zwłaszcza film 1 i 5 były zauważalnie blisko ocenianie. Zaufanie do filmowego przekazu prezentowanego na nagraniu 6 oceniony został zauważalnie blisko jak ten z nagrań 2 i 3, co widać zwłaszcza po medianie.

Na wstępie niniejszego podrozdziału słusznie założono, iż najwyższy wynik pod względem zaufania do filmowego przekazu wystąpi w nagraniach 4 i 5. Wbrew założeniom, nagranie 1 ocenione zostało zbieżnie z nimi (nie różni się, $p > 0,05$), zwłaszcza w grupie, która nie rozpoznała fałszerstwa nagrania. Na podstawie porównania parami wnioskować można, iż nawet słabszej jakości deepfake jest w stanie osiągnąć nieznacznie wyższy poziom zaufania do filmowego przekazu niż wideo z nieznanymi osobami.

4.8 Wnioski

Bezpieczeństwo narodowe w obszarze informacyjnym opiera się w dużej części na samoświadomości obywateli i ich zdolności percepcji oraz rozumowania. W niniejszym rozdziale starano się zweryfikować jaki jest obecny stan odporności społecznej na fałszywe materiały audiowizualne oraz na jakim poziomie jest zdolność ich postrzegania. Pytania zadawane respondentom pod prezentowanymi im nagraniami, w zestawieniu ze sobą, miały udzielić komplementarnej odpowiedzi na obecne możliwości rozpoznania fałszu.

Pytanie badawcze, jakie sformułowane zostało na wstępie niniejszego rozdziału brzmi: „czy internauci rozróżniają materiały multimedialne prawdziwe od fałszywych?". Próbą odpowiedzi na to pytanie jest uprzednio sformułowana hipoteza: internauci częściowo rozróżniają materiały multimedialne prawdziwe od fałszywych, wytworzonych przy wykorzystaniu technologii deepfake, ale mogą mieć trudności z odróżnieniem prawdziwych materiałów od fałszywych, zwłaszcza, jeśli są one dobrze wykonane.

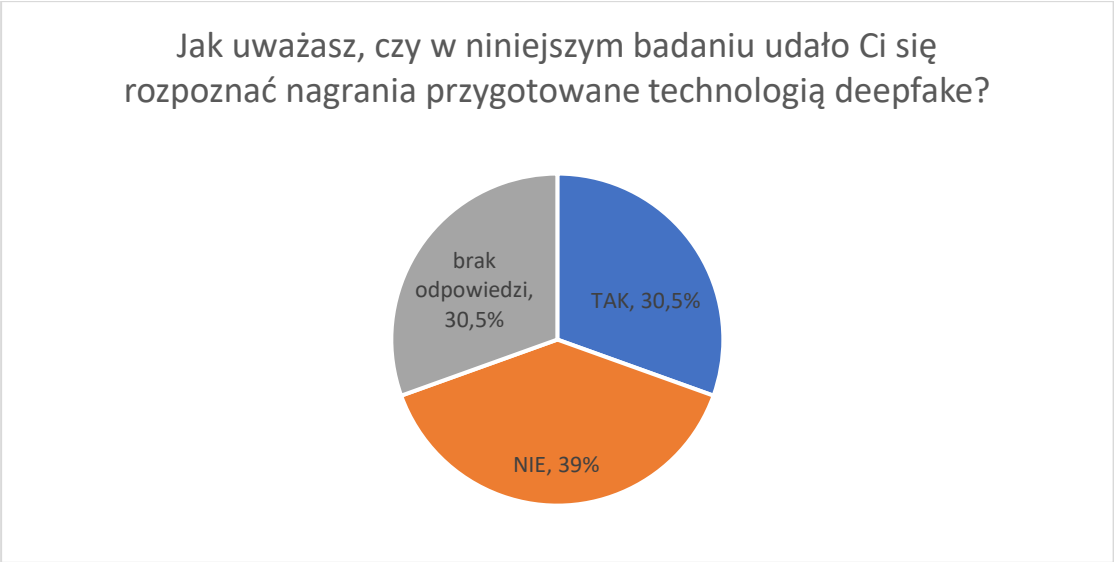
Hipotezę tę udało się zweryfikować pozytywnie na podstawie analizy materiału empirycznego zgromadzonego w formie odpowiedzi na sześć pytań zamkniętych oraz jedno pytanie otwarte (podrozdział 4.3). Analizie poddano odpowiedzi na pytania o naturalność nagrania, jego prawdziwość, rozpoznanie osób z nagrania, zaufanie do wizerunku osoby w nim występujących, kojarzenie tych osób oraz zaufanie do filmowego przekazu. Pomocna była w tym również analiza zgromadzonych danych jakościowych – odpowiedzi do pytania opisowego, dotyczącego elementów nienaturalnego wyglądu na nagraniach (artefaktów). Kompleksowa analiza pozwoliła zweryfikować stawianą hipotezę i odpowiedzieć na zadane pytanie badawcze.

Materiał empiryczny należy wpierw podzielić na dwa zbiory – elementów kluczowych oraz wspomagających. Pytania numer 11 i 13 choć podobne, nie są ze sobą tożsame. Oba miały charakter pytań dodatkowych, mających na celu wzmocnienie analizy odpowiedzi na pytania o zaufanie do osoby występującej na nagraniu oraz jej filmowego przekazu.

W toku analizy materiału empirycznego zgromadzonego w niniejszym rozdziale, zwłaszcza analizując odpowiedzi do pytań 1, 2, 3 i 10 stwierdzono, iż odpowiedzi na każde z czterech zamkniętych pytań w przypadku filmu pierwszego (deepfake) nie są istotnie różne od odpowiedzi dotyczących prawdziwego filmu czwartego, prezentującego średnio znanego influencera. Również w zestawieniu odpowiedzi z filmem piątym, na 3 z 4 pytań nie ma istotnej statystycznie różnicy pomiędzy grupami.

Drugi z filmów deepfake (nagranie 6) w większości przypadków jest natomiast statystycznie różny zarówno w zestawieniu z filmami z influencerami jak i w zestawieniu z pierwszym filmem deepfake. Był on bowiem gorszej jakości, a wytrenowanie go zajęło mniej czasu i uwagi twórcy niż film pierwszy. Skuteczność jego oddziaływania na próbę badawczą jest natomiast istotnie różna od filmów prezentujących nieznane osoby. W trakcie analizy ustalono jego naturalność jako pośredni pomiędzy filmami prezentującymi nieznane osoby, a tymi z średnio znanymi influencerami oraz pierwszym filmem deepfake. Potwierdza to również analiza odpowiedzi na pytanie opisowe. W przypadku tego filmu najwięcej osób zwraca uwagę na widoczne artefakty, wpływające na odbiór naturalności nagrania. Wykrycie oszustwa było w tym przypadku dużo łatwiejsze, co potwierdzają również osoby deklarujące nierozpoznanie deepfake. Może mieć na to wpływ znacznie mniejsza znajomość tej technologii w społeczeństwie w okresie prowadzenia eksperymentu (pierwsza połowa 2022 roku).

Analiza wewnątrzgrupowa, oparta na zmiennej nominalnej E (rozpoznanie nagrań deepfake) pozwoliła również ustalić, iż osoby deklarujące rozpoznanie fałszu odpowiadały statystycznie istotnie różnie na każde z analizowanych w niniejszym rozdziale pytań, od osób które zadeklarowały nierozpoznanie deepfake. Dzieje się tak zarówno dla pierwszego, jak i drugiego z nagrań deepfake. Potwierdza to prawdziwość odpowiedzi na pytanie E. Na pytanie „jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?" twierdzącej odpowiedzi udzieliło 30,5% badanych osób. 39% zadeklarowało, iż nie rozpoznało deepfake na żadnym z prezentowanych filmów, natomiast 30,5% nie udzieliło żadnej odpowiedzi.



Wykres 26 Wynik procentowy odpowiedzi na pytanie „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Na podstawie analizowanych odpowiedzi, stwierdzić można, iż w pewnej części internauci nie rozróżniają nagrań deepfake, a te prezentujące znane osoby wzbudzają w nich ufność. Ważnym elementem jest również grupa wiekowa badanych osób (od 18 do 33 lat), gdzie młodą grupę tę uznaje się za lepiej rozumiejącą „świat Internetu” oraz związane z nim zagrożenia od osób starszych. Zakłada się, że to właśnie grupa seniorów jest najbardziej podatna na manipulacje i oszustwa internetowe[247]. Osoby te są bardziej ufne i nie rozumieją technologii oraz zagrożeń obecnych w sieci. Osoby starsze są również bardziej skłonne do udzielania zaufania nieznajomym niż osoby młodsze. Sugeruje to, że osoby starsze mogą być bardziej podatne na manipulacje audiowizualną[248].

Sytuacja, w której internauta daje się zmanipulować oglądanemu materiałowi zdaje się być niezwykle ważna, zwłaszcza w modelu demokratycznym, gdzie obywatel poprzez narzędzia demokracji ma wpływ na władzę oraz społeczeństwo. Niepokojąca jest również sytuacja, w której obywatel nie jest pewien czy oglądany przez niego materiał jest prawdą czy kłamstwem. Powodować to może spadek zaufania do znajdywanych materiałów audiowizualnych oraz wzrost niepewności czy stanu napięcia. Największe konsekwencje niesie za sobą moment, w którym oglądający nagranie nie rozpozna fałszu i ulegnie jego przekazowi. Taka sytuacja staje się niezwykle niebezpieczna zwłaszcza

[247] M. Baran i inni, „Cybernauci – diagnoza wiedzy, umiejętności i kompetencji dzieci i młodzieży, rodziców i opiekunów oraz nauczycieli w zakresie bezpiecznego korzystania z Internetu. Raport podsumowujący badanie ex-ante”, Warszawa 2016.
[248] M. Horta, A. Shoenfelt, N. R. Lighthall, et al. „Age-group differences in trust-related decision-making and learning”. Sci Rep 14, 68 (2024).

w momencie pojawienia się dodatkowych elementów budujących narastanie niepewności czy strachu, takich jak konflikt zbrojny, stan podwyższonej gotowości czy ciężka sytuacja gospodarcza. Internauci częściowo rozróżniają materiały multimedialne prawdziwe od fałszywych, wytworzonych przy wykorzystaniu technologii deepfake, ale mogą mieć trudności z odróżnieniem prawdziwych materiałów od fałszywych, zwłaszcza, jeśli są one dobrze wykonane. Respondenci biorący udział w niniejszym badaniu, w większej części przekonani są co do prawdziwości nagrań, ponieważ nie dostrzegają minimalnie widocznych uchybień w obrazie. W zależności od przekonań mogą być w stanie uwierzyć w przekazywane im treści. To, że ludzie nie mają problem z rozpoznawaniem nagrań deepfake zdają się potwierdzać inne badania[249],[250].

W takim momencie, pojawienie się niezweryfikowanej, a przypuszczalnie prawdziwej informacji w formie nagrania deepfake prowadzić może do paniki społecznej, a być może wybuchu konfliktów. Trudno jest ocenić potencjalne skutki dobrze wygenerowanego filmu deepfake. Dotychczasowe próby, dzięki słabej jakości wideo oraz szybkiej reakcji mediów oraz rządów, kończyły się niepowodzeniem. Podręcznikowym przykładem może być wideo deepfake prezentujące Prezydenta Ukrainy poddającego państwo Rosji[251]. Dzięki szybkiej reakcji zarówno Prezydenta, jak i mediów oraz portali YouTube oraz Facebook i X (dawniej portal Twitter), udało się szybko usunąć spreparowane nagrania oraz dotrzeć do szerokiej grupy ludzi z przekazem demaskującym[252].

Wybuch wojny na Ukrainie spowodował w Polsce, przy pomocy plotki, nagły wzrost wypłat gotówki oraz wzrost osób tankujących pojazdy[253]. Po raz pierwszy w Polsce w wielu oddziałach oraz bankomatach pojawiły się deficyty gotówki. Skokowa liczba wypłat spowodowała miejscowo przejściowe problemy z dostępem do niej,

[249]Goh, D.HL., Lee, C.S., Chen, Z., Kuah, X.W., Pang, Y.L. (2022). „Understanding Users' Deepfake Video Verification Strategies". [w:] Stephanidis, C., Antona, M., Ntoa, S., Salvendy, G. (eds) HCI International 2022 – Late Breaking Posters. HCII 2022. Communications in Computer and Information Science, vol 1655. Springer, Cham.

[250] Domenteanu A, Tătaru G-C, Crăciun L, Molănescu A-G, Cotfas L-A, Delcea C. Living in the Age of Deepfakes: A Bibliometric Exploration of Trends, Challenges, and Detection Approaches. Information. 2024; 15(9):525.

[251] Fragment wideo wraz z komentarzem ekspertów, https://youtu.be/pfsdvbacYac [dostęp: 01.01.2023].

[252] Artykuł opisujący reakcję mediów społecznościowych na pojawienie się nagrania, https://edition.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html [dostęp: 01.01.2023].

[253] Artykuł opisujący wydarzenia z końca lutego 2022 roku, https://dziendobry.tvn.pl/newsy/problemy-z-wyplacaniem-gotowki-czy-pieniedzy-moze-zabraknac-5617004 [dostęp: 01.01.2023].

na skutek czego jeszcze większa liczba osób zechciała wypłacić swoje środki. Oficjalne komunikaty wydały najważniejsze w Polsce urzędy – Narodowy Bank Polski[254] oraz Komisja Nadzoru Finansowego[255].

Biorąc pod uwagę wyższą wiarygodność nagrań audiowizualnych, dobrze przygotowany materiał, umieszczony w odpowiednim czasie i odpowiednim miejscu, mógłby spowodować poważniejsze skutki. Takie nagranie może być częścią prowadzonej wojny informacyjnej, a towarzyszące jej publikacji działania mogą skutecznie zwiększyć lub osłabić jego działanie[256].

[254] Komunikat Narodowego Banku Polskiego zapewniający o bezpiecznym stanie gotówki, https://www.prawo.pl/podatki/wyplaty-gotowki-w-banku-i-z-bankomatu,513700.html [dostęp: 01.01.2023].
[255] Komunikat dotyczący braku planów w zakresie wprowadzania limitów wypłat, https://x.com/uknf/status/1496876307573096452 [dostep: 01.01.2023].
[256] D. Kazimierczak, „Walka informacyjna we współczesnych konfliktach i jej społeczne konsekwencje", Studia de Securitate et Educatione Civili 7, 2017.

Rozdział 5. Wpływ nagrań deepfake na bezpieczeństwo personalne w świetle przeprowadzonego eksperymentu

Od drugiej połowy 2021 roku, Internet został zdominowany przez nowy rodzaj oszustwa – fałszywe inwestycje. Zespół CSIRT KNF w okresie 01.01.2022 – 31.08.2022 wykrył i zgłosił 11 411 fałszywych reklam, prowadzących do oszustwa[257]. Za cały rok 2022 było to prawie 18 tysięcy reklam[258]. W reklamach często pojawiały się wizerunki znanych marek, takich jak Tesla, Orlen czy nazwy konkretnych banków obecnych na polskim rynku. Część z reklam prezentowała również bezprawnie wizerunki znanych osób, w tym celebrytów, sportowców, dyrektorów spółek oraz najważniejszych polityków. Często była to wyłącznie grafika prezentująca twarz danej osoby, czasami jednak, dla uwiarygodnienia swojego przekazu, oszuści przygotowywali krótkie, wycięte z kontekstu wideo. Z nagrań wynikało, iż osoba na nich prezentowana zachęca do powierzenia swoich pieniędzy oszustom.

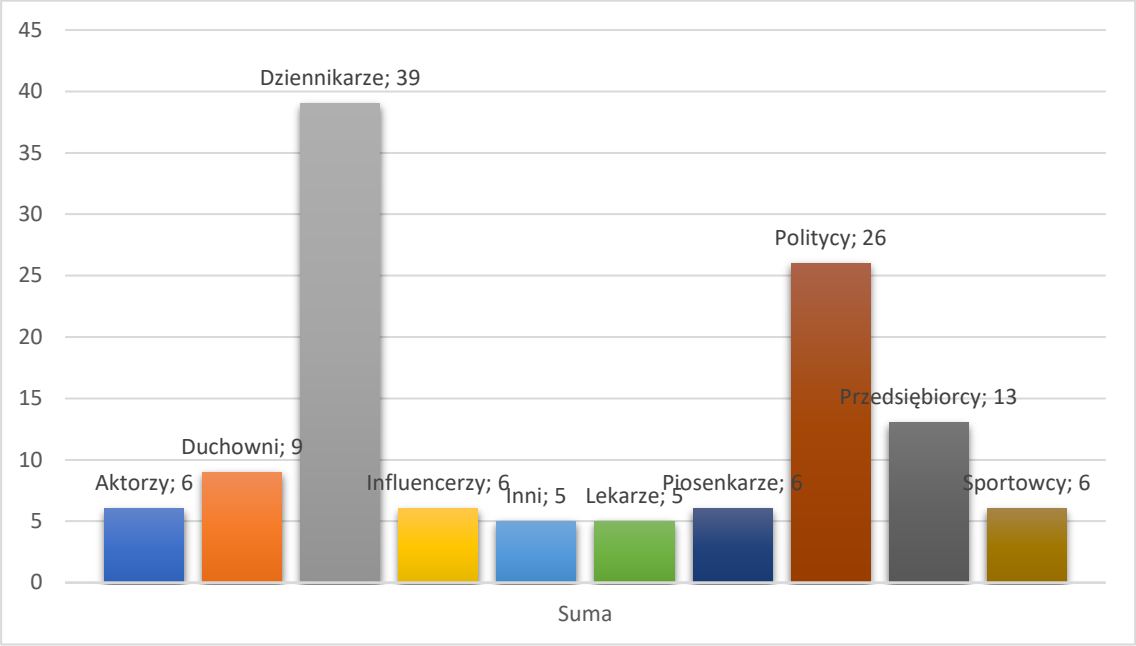
Wielokrotnie zdarzało się, iż były to nagrania polityków, którzy w specjalnie wyciętym i obrobionym fragmencie opowiadali o bezpieczeństwie i pewności danej inwestycji. Zarówno przed wyciętym fragmentem, jak i po, głos podkładał lektor, umiejscawiając dane przemówienie w stworzonej przez siebie rzeczywistości. Fragment wycięty był z szerszej wypowiedzi, w której polityk w żadnym stopniu nie odnosił się do owego programu inwestycyjnego. Całość brzmiała w ten sposób, iż osoba nieznająca zagrożeń, mogła odnieść wrażenie, że to faktycznie prezentowany polityk zachęca do inwestycji.

Osoby oszukane przy pomocy tej metody tracą zazwyczaj wszystkie zgromadzone przez siebie pieniądze. Często są to oszczędności ich całego życia[259]. Zdarza się również, iż zmanipulowane osoby zaciągają w bankach kredyty, po czym zmanipulowane, nieświadome konsekwencji, przekazują pieniądze oszustom.

W lipcu 2024 roku, zespół ekspertów z Pionu Sztucznej Inteligencji NASK we współpracy z zespołem CSIRT NASK, Pionem Ochrony Informacyjnej Cyberprzestrzeni

[257] Raport CSIRT KNF – Fałszywe Inwestycje, https://cebrf.knf.gov.pl/images/Raporty/Faszywe_inwestycje_2022.pdf [dostęp: 01.01.2023].
[258] Raport Roczny CSIRT KNF 2022.
[259] Artykuł opisujący jeden z przypadków, w którym emerytka straciła wszystkie swoje oszczędności, https://cebrf.knf.gov.pl/komunikaty/artykuly-csirt-knf/362-ostrzezenia/852-66-latka-stracila-prawie-190-tysiecy-zlotych-inwestujac-na-falszywej-platformie [dostęp: 01.01.2023].

oraz CSIRT KNF opublikował raport przedstawiający 121 nazwisk, które wykorzystywali oszuści do omawianego przestępstwa w 2024 roku, z wykorzystaniem technologii deepfake. Nazwiska zostały pogrupowane z kategoryzacją zawodu wykonywanego przez daną osobę. W raporcie określono 9 kategorii, takich jak dziennikarze, politycy, sportowcy, aktorzy, duchowni, influencerzy, lekarze, piosenkarze, przedsiębiorcy oraz pozostałe[260]. Poniżej zamieszczono tabelę wskazującą na dominujące grupy zawodowe wykorzystywane przez oszustów.



Wykres 27 Zsumowana ilość osób w danej kategorii zawodowej. Źródło: Raport NASK.

Z opublikowanego raportu wynika, iż aż 32% wykorzystanych w oszustwie wizerunków należało do dziennikarzy, a 21% do polityków. Oszukańcze nagrania publikowane były na platformach społecznościowych grupy Meta – Facebook oraz instagram, TikTok, X (dawniej Twitter) oraz YouTube.

W 2022 roku, powstały w krajach anglojęzycznych pierwsze oszustwa wykorzystujące technologię deepfake do przekonywania do inwestycji według zaprezentowanego powyżej schematu[261]. Oszuści przy pomocy deepfake tworzą fałszywe nagrania, na których znana osoba reklamuje oszukańczą inwestycję. Wraz

[260] Raport NASK, „Lista osób, których wizerunek przestępcy wykorzystali w oszustwach deepfake", https://www.nask.pl/pl/aktualnosci/5429,NASK-ostrzega-przestepcy-tworza-coraz-bardziej-pomyslowe-oszustwa-deepfake-Bezpr.html [dostęp: 01.08.2024].
[261] Artykuł opisujący, jak oszuści utworzyli fałszywe nagranie Elona Muska i zachęcali przy jego pomocy do inwestowania na nieistniejącej giełdzie, https://www.outlookindia.com/business/criminals-use-elon-musk-s-deepfake-video-to-dupe-crypto-investors-crypto-market-rises-news-198403 [dostęp: 01.01.2023].

z rozwojem technologii zakłada się, iż takie scenariusze będą coraz bardziej realne, a prezentowane nagrania będą coraz doskonalsze[262].

W połowie 2023 roku również w Polsce, zaczęły pojawiać się pierwsze oszukańcze nagrania deepfake, na których głos i usta danych polityków obrabiane były w taki sposób, by oglądający nagranie mógł odnieść wrażenie, iż to faktycznie dany polityk dane sformułowanie wypowiedział. Przy pomocy aplikacji pozwalających podrabiać dźwięk, oszuści obrabiali głos danej osoby, dostosowując do swojego tekstu, zaś dzięki aplikacji do edycji deepfake wideo, byli w stanie dostosować mimikę twarzy do fałszywej wypowiedzi. Tak wyglądające nagrania zamieszczane są w mediach społecznościowych i powodują, iż oglądający je mogą nie być w stanie rozpoznać fałszu[263].

Nagrania deepfake mogą nie tylko zachęcać do inwestowania na danej platformie, lecz również manipulować giełdą i wpływać na cenę akcji[264]. Oszuści mogą wytworzyć fałszywe nagranie znanego inwestora, prezesa firmy lub polityka, na którym będzie on zachęcał lub zniechęcał do inwestowania. Osoba z nagrania może również prezentować wymyślone przez siebie, istotne dla wartości spółki informacje, tym samym wpływając na wartość swojego portfela. Istnieje ryzyko, iż w najbliższym czasie powstawać będą coraz częściej oszukańcze nagrania, na których znane osoby zachęcać będą do inwestowania w konkretną spółkę lub informować o jej problemach i zachęcać do szybkiej sprzedaży akcji.

Oba scenariusze opierają się na tworzeniu fałszywych nagrań, złożonych z fałszywego obrazu, a coraz częściej również i dźwięku. By przeanalizować obecne możliwości technologii deepfake oraz jej wpływie na decyzje badanych, w ankiecie umieszczono szereg pytań dotyczących odbioru wideo oraz odczuciach osób je oglądających.

Pytanie badawcze, postawione w niniejszym rozdziale brzmi: jaki wpływ ma percepcja zmanipulowanych nagrań wideo na wybory i opinie internautów, w świetle

[262] Artykuł opisujący próby podszywania się pod znanych inwestorów, https://cointelegraph.com/news/sam-bankman-fried-deepfake-attempts-to-scam-investors-impacted-by-ftx [dostęp: 01.01.2023].
[263] Przykłady fałszywych nagrań deepfake, zachęcających do fałszywej inwestycji dostępne są na stronie Centrum Edukacji dla Bezpieczeństwa Rynku Finansowego, https://cebrf.knf.gov.pl/deepfake [dostęp: 01.02.2024].
[264] Artykuł opisujący możliwości wykorzystania deepfake na cenę akcji, https://cointelegraph.com/news/here-s-how-to-quickly-spot-a-deepfake-crypto-scam-cybersecurity-execs [dostęp: 01.01.2023].

bezpieczeństwa narodowego? Weryfikowana hipoteza brzmi: percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów, a jej oddziaływanie na bezpieczeństwo narodowe jest zależne od stopnia wiarygodności tych nagrań oraz od kontekstu, w jakim są one prezentowane. Celem odpowiedzi na postawione pytanie badawcze i weryfikacji hipotezy, zdecydowano się na analizę materiału empirycznego złożonego z odpowiedzi na sześć pytań. Odpowiedzi na każde z nich przeanalizowane zostały w osobnym podrozdziale. Siódmy podrozdział zawiera w sobie syntetyczną analizę prezentowanych we wcześniejszych podrozdziałach wnioskach.

Pierwszy podrozdział zawiera zagadnienie zachęcania fałszywymi nagraniami do inwestycji. Na podstawie analizy odpowiedzi na zadane pytanie, przebadane zostało czy respondenci różnie oceniali wpływ prezentowanych nagrań na chęć inwestowania. w przypadku statystycznych różnic, mogłoby to wskazywać, iż nagrania deepfake były w mniejszym lub większym stopniu skuteczne od nagrań średnio znanych influencerów lub nieznanych osób.

W następnym podrozdziale zaprezentowano analizę odpowiedzi na pytanie dotyczące tego, ile osoba oglądająca nagranie byłaby skłonna zainwestować na prezentowanej platformie. Zastanawiające jest, czy skuteczność reklamy przedstawianej na poszczególnych nagraniach w różny sposób wpływała na procent oszczędności, które oglądający je może przekazać. Pytanie to miało na celu potwierdzenie w jak znacznym stopniu zaufanie do nagrania przekłada się na chęć podjęcia ryzyka i powierzenia własnych pieniędzy.

Trzeci podrozdział opisano analizę odpowiedzi respondentów dotyczące ich przekonania o realności możliwości osiągnięcia dużego zysku. Pytanie miało na celu ustalenie, w jakim stopniu widzowie nagrań wierzą w przedstawione w nich deklaracje, które obiecują wysoką stopę zwrotu bez ryzyka inwestycyjnego. Wyniki tej analizy mogą wskazać, w jakim stopniu nagrania średnio znanych influencerów oraz deepfake różnią się od nagrań osób nieznanych pod względem skuteczności w przekonywaniu do nierealistycznych obietnic.

Czwarty podrozdział opisuje analizę obawy respondentów dotyczące utraty pieniędzy. Wszystkie sześć nagrań dotyczyło fałszywych inwestycji, w których zaangażowanie skutkowałoby całkowitą utratą środków. Zastanawiające jest, jak różni

się podatność widzów na przekaz zawarty w nagraniach w zależności od filmu oraz jak poszczególne filmy wpływają na poziom obaw wśród osób je oglądających.

Piąty podrozdział to analiza na ile respondenci obawiają się, iż dane wideo wpłynie na decyzje inwestycyjne innych osób. Zastanawiające jest, w jakim stopniu oglądający nagranie oceniają kompetencje cyfrowe społeczeństwa oraz które nagrania ich zdaniem mogą wywrzeć na społeczeństwo największy wpływ.

W szóstym podrozdziale zamieszczona została analiza odpowiedzi na pytanie zbliżone, choć nie tożsame w swojej treści do pytania analizowanego w piątym rozdziale. Tym razem respondenci spytani zostali jednak o to jak ich zdaniem prezentowane nagranie może wpłynąć na chęć zainwestowania w dany projekt. Badana jest więc opinia na potencjalny wpływ poszczególnych filmów na społeczeństwo.

Siódmy podrozdział zawiera prezentację wniosków w odniesieniu do pytania badawczego postawionego w niniejszym rozdziale. Weryfikacja hipotezy odbyła się na podstawie opisanej we wcześniejszych podrozdziałach analizy materiału empirycznego.

5.1 Zdolność do przekonywania nagrań deepfake

Pierwszym z pytań w ankiecie, które bezpośrednio dotyczyło oszustw inwestycyjnych było pytanie dotyczące zachęcenia do inwestycji. Na celu miało wykrycie czy respondenci są skłonni zainwestować w reklamowaną przez osoby występujące w filmach inwestycję. Pytanie brzmiało: „w jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?”. Respondenci podobnie jak w przypadku pozostałych pytań, swe odpowiedzi udzielali na dziesięciostopniowej skali, gdzie 1 oznaczało „w ogóle”, zaś 10 „bardzo”.

Filmy wyświetlano w losowej kolejności. Nagranie pierwsze oraz szóste w niniejszej analizie były filmami deepfake, gdzie na twarze zatrudnionych osób nałożono twarze dwójki influencerów, przy czym wideo pierwsze prezentowało kobietę, bardziej rozpoznawalną niż mężczyznę z nagrania szóstego. Filmy drugi i trzeci w niniejszej analizie, to nagrania prezentujące nieznane, obce osoby, zachęcające do inwestycji. Filmy czwarty i piąty to prawdziwe nagrania średnio znanych influencerów, również zachęcających do tej samej inwestycji. Teksty wypowiadane na nagraniach były ze sobą zbieżne dla każdego filmu i opisane zostały w rozdziale 1.2.2.

Dla niniejszego pytania, założono, iż najwyższą ocenę otrzymają nagrania 4 i 5, ze względu na zaufanie do popularnych osób, a także lepszą jakość nagrań. Założono również, iż chęć inwestycji, wśród osób które nie rozpoznały deepfake będzie na zbliżonym do nagrań 4 i 5 poziomie. Nagrania 2 i 3 ze względu na słabą jakość oraz nieznane twarze powinny otrzymać najniższe oceny. Poniżej zaprezentowano statystki opisowe dla wszystkich nagrań.

Statystyki opisowe

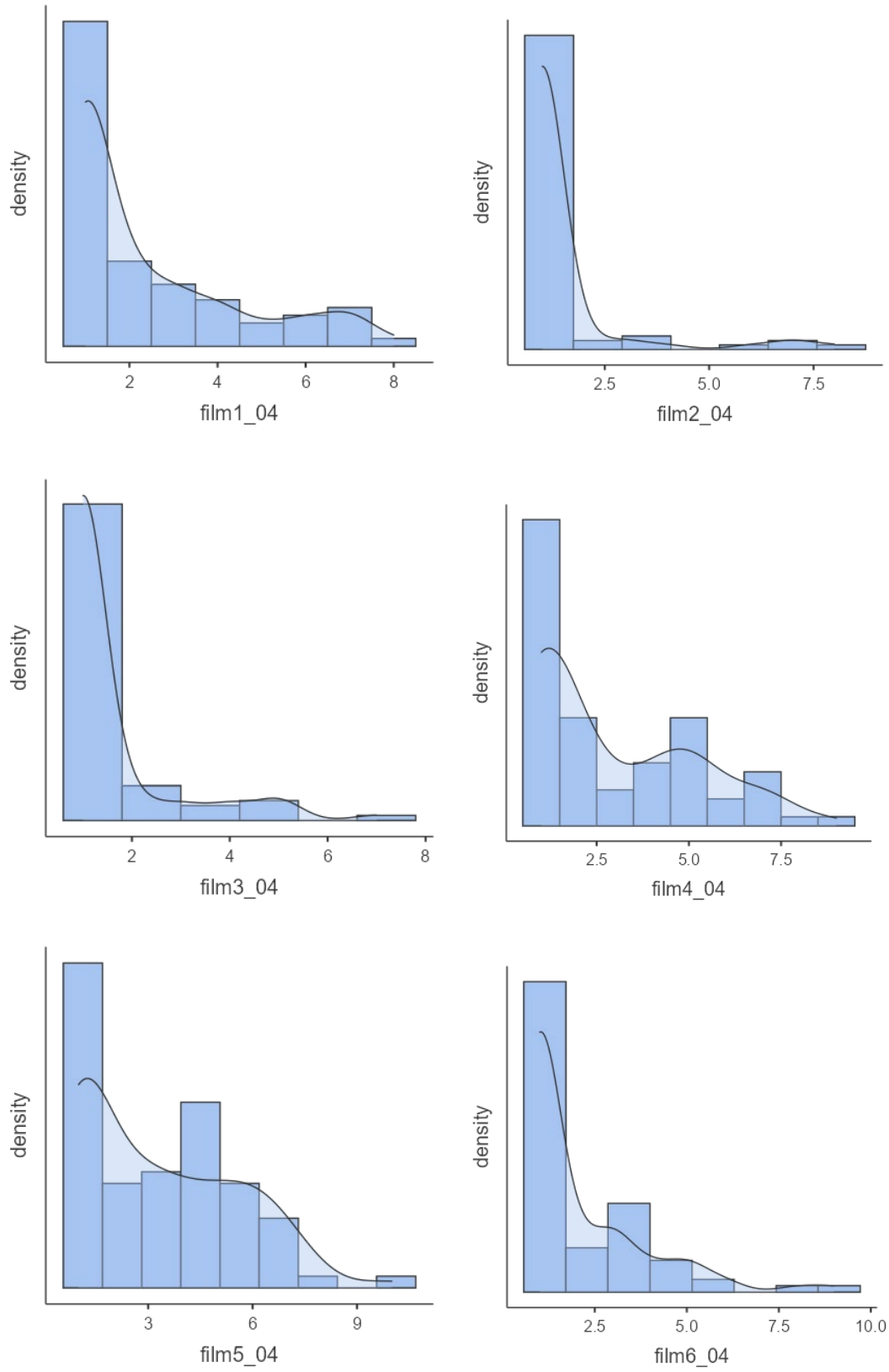
	film1_04	**film2_04**	**film3_04**	**film4_04**	**film5_04**	**film6_04**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	2.42	1.41	1.52	2.94	3.27	2.06
SE	0.222	0.155	0.141	0.247	0.253	0.196
95% CI dolna granica przedziału ufności dla średniej	1.99	1.11	1.24	2.45	2.78	1.68
95% CI górna granica przedziału ufności dla średniej	2.86	1.72	1.79	3.42	3.77	2.45
Me	1.00	1.00	1.00	2.00	3.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.99	1.38	1.25	2.21	2.27	1.74
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	8.00	8.00	7.00	9.00	10.0	9.00
SKE	1.31	3.68	2.57	0.836	0.666	1.92
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	0.530	13.0	6.11	-0.484	-0.509	3.70
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.741	0.335	0.482	0.820	0.871	0.675
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 25 Statystyki opisowe dla czwartego pytania „W jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?”.

Tabela 17 przedstawia statystki opisowe dotyczące odpowiedzi na czwarte pytanie: „w jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?”. W kolumnach znajdują się statystyki dotyczące odpowiedzi na to pytanie dla każdego z nagrań. Film pierwszy przedstawiał wideo deepfake z nałożoną twarzą znanej aktorki zachęcającej do inwestycji na fałszywej platformie inwestycyjnej. Wideo drugie i trzecie przedstawiało nieznane, obce osoby, które również zachęcały do tej samej inwestycji.

Nagrania czwarte i piąte przedstawiały średnio popularnych influencerów, którzy również zachęcali do zainwestowania na fałszywej platformie. Wideo szóste, podobnie jak pierwsze, było nagraniem deepfake i prezentowało średnio znanego influencera reklamującego tę samą inwestycję. Podkreślić należy, iż w przypadku zainwestowania swoich pieniędzy na tej platformie, osoba mogła stracić wszystkie swoje oszczędności. Na podstawie przeanalizowanych kazusów opartych na tym schemacie oszustwa, często poszkodowani oprócz swoich oszczędności zaciągają również zobowiązania finansowe (kredyty), z których pozyskane środki również przekazują oszustom.

Dane z powyższej tabeli pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego (dla odpowiedzi na czwarte pytanie – w przypadku każdego filmu). Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$ (najwyższy raportowany – w większości nauk społecznych – poziom istotności statystycznej), wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności (jeden ze stosowanych wyznaczników braku rozkładu normalnego), a w większości przypadków, są one większe od 1 oraz wartości bezwzględne kurtozy, które w połowie przypadków są większe od wartości błędu standardowego tej miary (kolejny ze stosowanych wskaźników braku rozkładu normalnego), zaś w pozostałych przypadkach różnią się od niej nieznacznie. Z uwagi na brak rozkładu normalnego uzyskanych wyników, w dalszych analizach statystycznych zastosowano testy nieparametryczne. Testy parametryczne, które opierają się na średnich, można stosować jedynie w przypadku, gdy rozkład wyników jest zbliżony do normalnego (ich niewłaściwe użycie w sytuacji braku normalności może prowadzić do błędnych wyników oraz mylnych wniosków). Poniższe wykresy przedstawiają histogramy (rozkłady) uzyskanych wyników dla każdego filmu, dotyczące czwartego pytania: „W jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?”. Histogramy te także ilustrują, że żaden z rozkładów wyników nie jest zbliżony do rozkładu normalnego.



Wykres 28 Zbiór 6 wykresów odpowiedzi na czwarte pytanie dla każdego z sześciu filmów.

Na powyższych wykresach widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują silną prawoskośność dla rozkładów, ze znaczną dominacją odpowiedzi 1

– wcale. Dzieje się tak zwłaszcza dla filmu 2 i 3 (nieznane osoby). Niewiele lepiej oceniany jest film 6 (deepfake). Filmy 1 (deepfake) i 4 (średnio znany influencer) są do siebie zbliżone pod względem rozkładu. Najwyższe wartości zdaje się mieć film 5 (średnio znany influencer), co sugerować może o jego najwyższej skuteczności pod względem zachęcania do inwestycji.

Celem weryfikacji czy odpowiedzi na pytanie dotyczące zachęcenia do inwestycji różne były w zależności od tego czy respondent rozpoznał czy nie rozpoznał nagrania deepfake, zdecydowano się na dalszą analizę z uwzględnieniem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Konieczne było bowiem sprawdzenie czy osoby, które nie rozpoznały fałszu nagrań 1 i 6 (deepfake) były przekonane do prawdziwości inwestycji w podobnym stopniu jak po obejrzeniu pozostałych nagrań oraz czy nie zaobserwowano różnic w skłonności do inwestowania pomiędzy filmami deepfake, a prawdziwymi, z uwzględnieniem tego czy ktoś rozpoznał czy nie rozpoznał fałszu na nagraniach.

Statystyki opisowe

	zmienna nominalna E	**film1_04**	**film2_04**	**film3_04**	**film4_04**	**film5_04**	**film6_04**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	3.16	1.41	1.31	3.03	3.16	2.75
	tak	1.80	1.48	1.56	2.72	3.44	1.36
SE	nie	0.404	0.215	0.145	0.398	0.387	0.378
	tak	0.337	0.337	0.317	0.434	0.504	0.207
95% CI dolna granica przedziału ufności dla średniej	nie	2.36	0.985	1.03	2.25	2.40	2.01
	tak	1.14	0.819	0.939	1.87	2.45	0.954
95% CI górna granica przedziału ufności dla średniej	nie	3.95	1.83	1.60	3.81	3.91	3.49
	tak	2.46	2.14	2.18	3.57	4.43	1.77
Me	nie	2.50	1.00	1.00	2.00	3.00	2.00

Statystyki opisowe

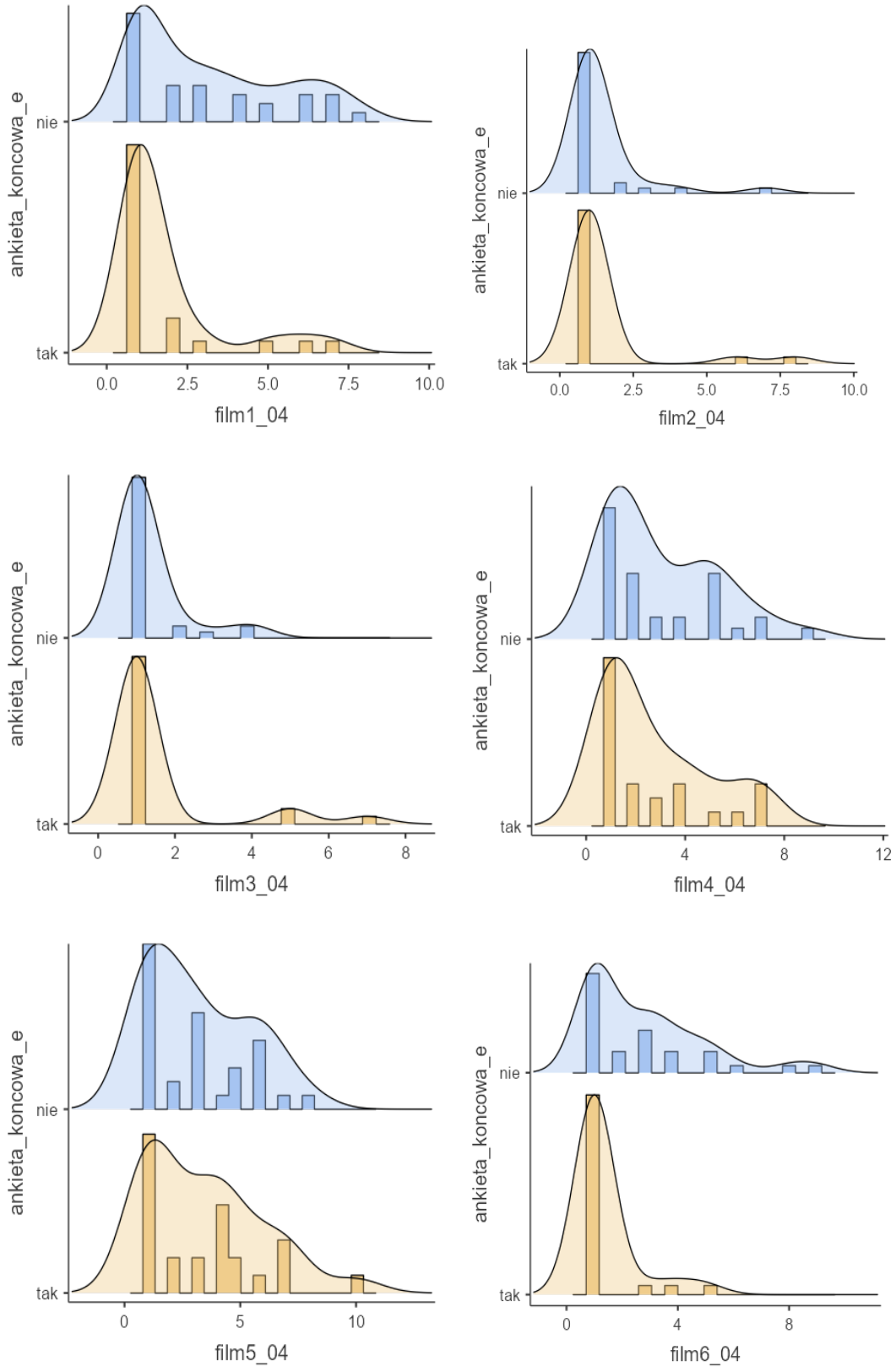
	zmienna nominalna E	film1_04	film2_04	film3_04	film4_04	film5_04	film6_04
	tak	1.00	1.00	1.00	2.00	3.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.29	1.21	0.821	2.25	2.19	2.14
	tak	1.68	1.69	1.58	2.17	2.52	1.04
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	8.00	7.00	4.00	9.00	8.00	9.00
	tak	7.00	8.00	7.00	7.00	10.0	5.00
SKE	nie	0.692	3.75	2.71	0.919	0.597	1.36
	tak	2.28	3.47	2.72	1.00	0.868	2.85
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.893	15.3	6.46	-0.0365	-0.926	1.56
	tak	4.30	11.3	6.45	-0.367	0.203	7.35
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.846	0.393	0.436	0.838	0.854	0.805
	tak	0.551	0.316	0.401	0.775	0.867	0.400
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	0.004	< .001

Tabela 26 Statystyki opisowe dla czwartego pytania „W jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Powyższa tabela przedstawia analogiczne dane jak poprzednia tabela, natomiast w tym przypadku rozkłady zmiennych zostały podzielone względem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości bezwzględne kurtozy, gdzie poza nagraniem czwartym, są one wyższe od błędu standardowego tej miary. Kolejnym wyznacznikiem jest wartość skośności, gdzie w każdym przypadku wartość ta była większa od błędu standardowego skośności. O braku rozkładu normalnego świadczą również niskie wyniki ($p < 0{,}001$ dla większości nagrań, poza filmem piątym, a $p < 0{,}05$ dla wszystkich) testu Shapiro-Wilk.

Na poniższych wykresach przedstawiono histogramy dla każdego filmu, z uwzględnieniem zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”).



Wykres 29 Zbiór 6 wykresów odpowiedzi na czwarte pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na czwarte pytanie, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 6 (deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi (przewaga oceny 1) dla grupy, która rozpoznała, iż jest to nagranie fałszywe. Zastanawiająca może być korelacja wyników odpowiedzi dla filmu pierwszego z pozostałymi, gdyż ich układ zdaje się być podobny do odpowiedzi dla filmu 4 i 5.

W celu oceny czy różnice te są istotne statystycznie, przeprowadzono szereg nieparametrycznych testów Kruskal-Wallis, które stanowią alternatywę dla jednoczynnikowej analizy wariancji (ANOVA). Wyniki tych testów znajdują się w poniższej tabeli. Zauważono, że jedynie w przypadku pierwszego i ostatniego filmu (oba deepfake) rozkłady różnią się istotnie ($p < 0,05$). To sugeruje, że różnice między tymi dwiema grupami są na tyle znaczące, że można je analizować za pomocą metod statystycznych. Istotność statystyczna została wykorzystana do stwierdzenia czy różnice między grupami są wystarczająco duże, aby uznać je za realne i nieprzypadkowe.

Kruskal-Wallis

	χ^2	df	p	ε^2
Film1_04	7.0217	1	0.008	0.12539
film2_04	0.5587	1	0.455	0.00998
film3_04	0.0347	1	0.852	6.21e-4
film4_04	0.4175	1	0.518	0.00746
film5_04	0.1153	1	0.734	0.00206
film6_04	10.5017	1	0.001	0.18753

Tabela 27 Test Kruskal-Wallis dla odpowiedzi do pytania czwartego.

Na podstawie powyższej tabeli, wnioskować można, iż respondenci deklarujący rozpoznanie fałszywych nagrań odpowiadali na to pytanie w odmienny sposób niż osoby, które fałszu nie rozpoznały (dla nagrań deepfake). Istotność statystyczna stwierdzona w przypadku filmu pierwszego i szóstego wskazuje, iż respondenci rozpoznający deepfake nie byli skłonni do inwestycji w takim stopniu jak osoby nierozpoznające fałszu.

W celu weryfikacji hipotezy dotyczącej różnic w odpowiedziach na pytanie czwarte („W jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?”)

w kontekście różnych filmów, zastosowano nieparametryczny test Friedmana, który stanowi odpowiednik analizy wariancji z powtarzalnymi pomiarami (RM ANOVA). Wyniki analizy wskazały, że odpowiedzi na to pytanie różnią się istotnie pomiędzy wszystkimi filmami (prawdziwymi i fałszywymi) z wartością χ^2 (5) = 101; p < 0,001. W celu dalszej analizy różnic między odpowiedziami, przeprowadzono test Durbin-Conover jako nieparametryczny odpowiednik testów post hoc. Wyniki znajdują się w poniższej tabeli.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_04	-	film2_04	5.083	< .001
film1_04	-	film3_04	4.765	< .001
film1_04	-	film4_04	2.255	0.025
film1_04	-	film5_04	3.939	< .001
film1_04	-	film6_04	1.493	0.136
film2_04	-	film3_04	0.318	0.751
film2_04	-	film4_04	7.338	< .001
film2_04	-	film5_04	9.022	< .001
film2_04	-	film6_04	3.590	< .001
film3_04	-	film4_04	7.020	< .001
film3_04	-	film5_04	8.704	< .001
film3_04	-	film6_04	3.272	0.001
film4_04	-	film5_04	1.684	0.093
film4_04	-	film6_04	3.748	< .001
film5_04	-	film6_04	5.432	< .001

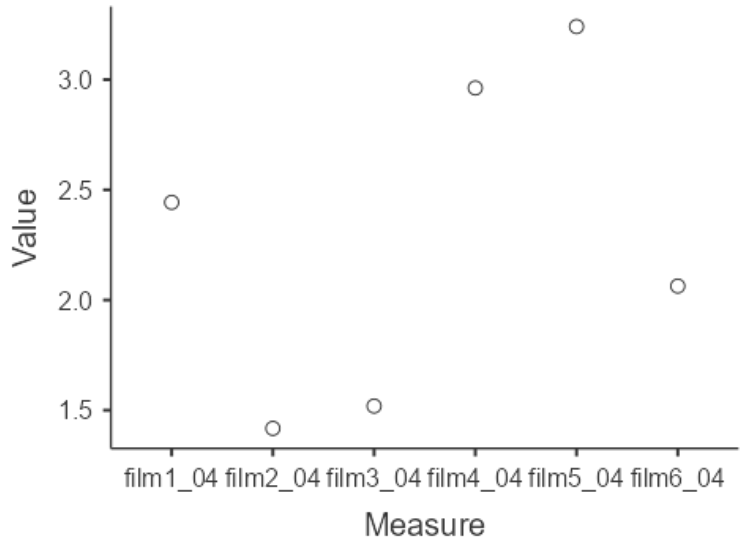
Tabela 28 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania czwartego.

Powyższe obliczenia pokazały, że przekaz skłonności do inwestycji w filmie 1 (deepfake) różni się od filmu 2, 3, 4 i 5 na co wskazuje p < 0,05, natomiast nie zaobserwowano różnicy pomiędzy nim a przekazem w filmie 6 (p > 0,05) (również nagranie deepfake). Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4, 5 oraz 6. Respondenci w większym stopniu zostali zachęceni do inwestycji przez znane osoby, niż przez osoby zupełnie nieznane. Pomiędzy filmem 4 i 5 nie zaobserwowano różnic. Z porównania Durbin-Conover wiemy, że pomiędzy drugim nagraniem deepfake – filmem 6 – a 4 i 5 (prawdziwe nagrania influencerów) zaobserwowano różnice.

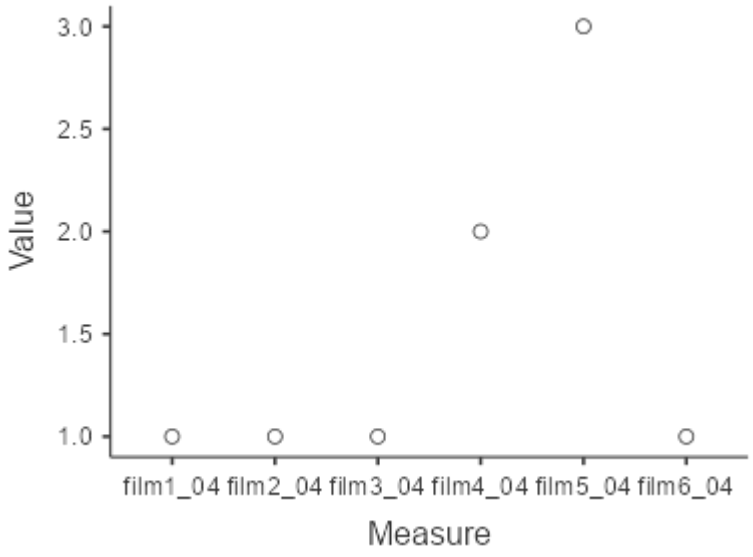
Wnioskować można, iż zachęcenie do inwestycji w filmach deepfake 1 i 6 było istotnie różne niż w filmach prawdziwych influencerów. Jednocześnie nagrania deepfake

były istotnie różne od nagrań prezentujących nieznane osoby. Wnioskować można, iż oba nagrania deepfake w podobnym stopniu wpłynęły na skłonność do inwestycji respondentów.

W celu weryfikacji tej hipotezy, poniżej umieszczono wykresy prezentujące średnie oraz mediany dla każdego z filmów. Jak zauważono, rozkłady wyników są istotnie różne od normalnego, co czyni średnie mało wiarygodnymi. W związku z tym zaleca się korzystać z mediany, która w przeciwieństwie do średniej nie jest obciążoną miarą tendencji centralnej.



Wykres 30 Średnia odpowiedzi dla pytania czwartego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 31 Mediana odpowiedzi dla pytania czwartego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Z prezentowanych powyżej wykresów odczytać można, iż średnia odpowiedzi dla filmu 4 i 5 jest do siebie zbliżona. Nieznacznie niżej ocenione zostało zachęcenie do inwestycji dla filmów deepfake, zwłaszcza pierwszego. Z przeprowadzonych testów Durbin-Conover, wiadomo jednak, iż ocena zachęcenie do inwestycji dla każdego z filmów deepfake była istotnie różna względem filmów prezentujących influencerów. Jednak biorąc pod uwagę grupę osób, które twierdzą, iż nie rozpoznały nagrań deepfake, skuteczność filmów deepfake oceniania jest przez respondentów na tym samym poziomie (mediana deepfake 1,5 i 2, zaś filmów z influencami 2 oraz 3).

Najniżej ocenione zostały filmy nieznanych osób, w przypadku których średnia (1,41 i 1,52) była bliska oceny jeden (w ogóle). Nagrania deepfake w porównaniu do tych filmów ocenione zostały istotnie różnie, co wykazuje również średnia wyników. Mediana dla filmów deepfake jak i osób nieznanych jest na tym samym poziomie (Me = 1). Wnioskować można, iż o ile skuteczność w przekonywaniu do fałszywych inwestycji nagrań deepfake, nie była na tyle skuteczna co filmów prawdziwych influencerów, to nadal ocena ta była wyższa niż w przypadku prezentowania nagrań twarzy nieznanych osób.

5.2 Skłonność do inwestowania

Do prawidłowej oceny skuteczności przygotowanych filmów, postanowiono odpytać respondentów o procent oszczędności jaki skłonni by byli przekazać

na reklamowanej na nagraniach platformie inwestycyjnej. Zadano pytanie „jaki procent swoich oszczędności byłbyś skłonny zainwestować na polecanej platformie po obejrzeniu tego nagrania?”. Odpowiedź udzielano na dziesięciostopniowej skali, gdzie 1 to w ogóle, 10 wszystkie.

Założono, iż respondenci najwięcej będą chcieli zainwestować po zachęcie średnio znanych influencerów. Oczekiwano, iż wyniki filmów deepfake będą zbliżone do tych osiągniętych w filmach 4 i 5 oraz będą statystycznie różne od nagrań osób nieznanych (filmy 2 i 3). Poniżej prezentowane są statystyki opisowe dla wszystkich filmów.

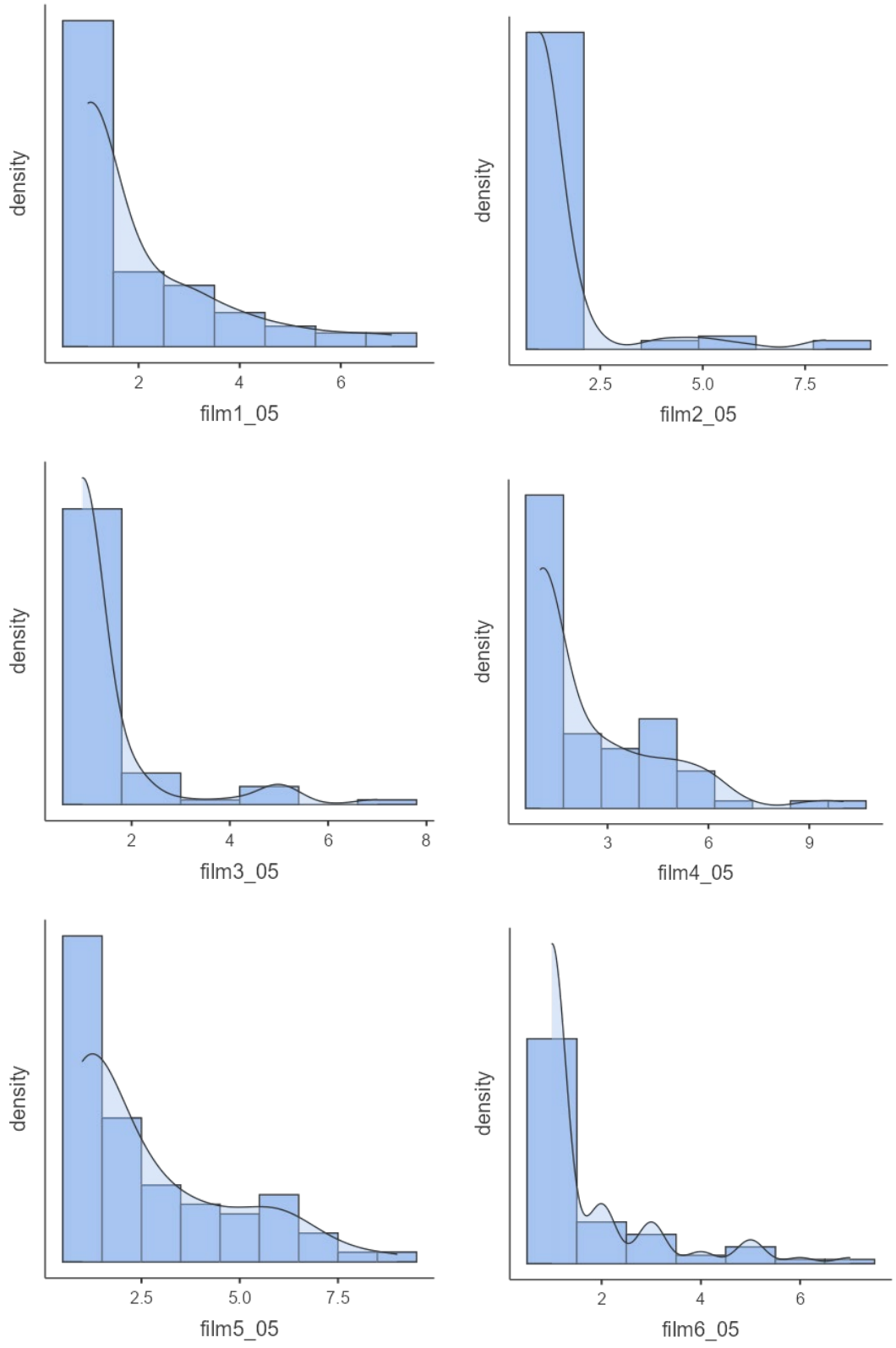
Statystyki opisowe

	film1_05	**film2_05**	**film3_05**	**film4_05**	**film5_05**	**film6_05**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	1.98	1.46	1.42	2.45	2.71	1.72
SE	0.172	0.159	0.131	0.228	0.233	0.151
95% CI dolna granica przedziału ufności dla średniej	1.64	1.15	1.16	2.00	2.26	1.43
95% CI górna granica przedziału ufności dla średniej	2.31	1.77	1.67	2.90	3.17	2.02
Me	1.00	1.00	1.00	1.00	2.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.53	1.42	1.16	2.04	2.08	1.34
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	7.00	8.00	7.00	10.0	9.00	7.00
SKE	1.70	3.46	3.15	1.56	1.12	2.11
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	2.28	11.8	9.69	2.20	0.246	4.03
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.691	0.372	0.415	0.746	0.804	0.612
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 29 Statystyki opisowe dla piątego pytania „Jaki procent swoich oszczędności byłbyś skłonny zainwestować na polecanej platformie po obejrzeniu tego nagrania?”.

Analiza danych z powyższej tabeli wyraźnie wskazuje, że nie możemy założyć, iż rozkład uzyskanych wyników przypomina rozkład normalny. To stwierdzenie znajduje potwierdzenie w wartościach testu Shapiro-Wilk, które dla wszystkich przypadków mają $p < 0,001$. Dodatkowo, wartości skośności dla każdego filmu są wyższe od błędów standardowych skośności oraz przekraczają 1. Wartości kurtozy, z wyjątkiem filmu 5, również są większe niż odpowiadające im błędy standardowe. W związku z brakiem

rozkładu normalnego w analizowanych wynikach, w dalszych badaniach zastosowano testy nieparametryczne.



Wykres 32 Zbiór 6 wykresów odpowiedzi na piąte pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Respondenci nie są skorzy do potencjalnego inwestowania swoich oszczędności. Dzieje się tak zarówno po obejrzeniu jednego z nagrań deepfake (szósty film), jak i osób nieznanych (nagranie drugie i trzecie). Dopiero średnio znani influencerzy oraz pierwszy film deepfake powodują u części z osób chęć do inwestycji pewnego procenta swoich oszczędności. Influencer występujący na 5 filmie zdaje się w największym stopniu przekonywać do powierzenia swoich oszczędności na platformie inwestycyjnej. Drugi z średnio znanych influencerów, osiąga wyniki zbieżne z tymi uzyskanymi przez osobę występującą na pierwszym nagraniu deepfake.

Aby zweryfikować tę hipotezę, należy ustalić, czy zdolność do rozpoznania fałszu wpłynęła na postrzeganą wiarygodność osób prezentowanych w nagraniach. Poniższa tabela zawiera dane porównywalne z poprzednią, przy czym w tym przypadku rozkłady zmiennych zostały podzielone według zmiennej nominalnej E, której treść brzmi: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Statystyki opisowe

	zmienna nominalna E	**film1_05**	**film2_05**	**film3_05**	**film4_05**	**film5_05**	**film6_05**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	2.47	1.31	1.34	2.88	2.47	2.22
	tak	1.56	1.68	1.48	2.08	2.96	1.20
SE	nie	0.327	0.176	0.199	0.448	0.327	0.294
	tak	0.271	0.399	0.265	0.336	0.495	0.129
95% CI dolna granica przedziału ufności dla średniej	nie	1.83	0.967	0.954	2.00	1.83	1.64
	tak	1.03	0.898	0.960	1.42	1.99	0.947
95% CI górna granica przedziału ufności dla średniej	nie	3.11	1.66	1.73	3.75	3.11	2.79
	tak	2.09	2.46	2.00	2.74	3.93	1.45

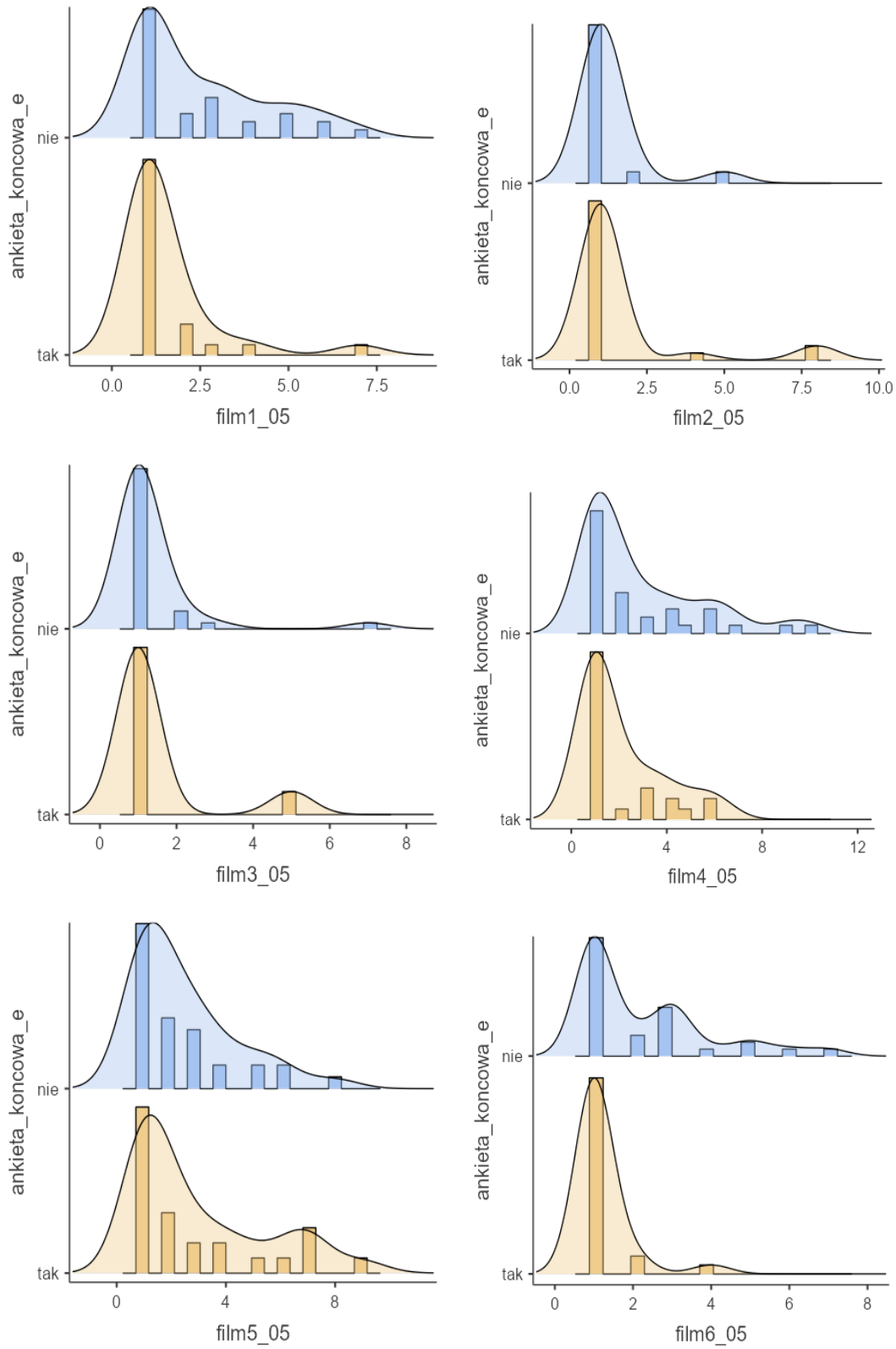
Statystyki opisowe

	zmienna nominalna E	film1_05	film2_05	film3_05	film4_05	film5_05	film6_05
Me	nie	1.50	1.00	1.00	2.00	2.00	1.00
	tak	1.00	1.00	1.00	1.00	2.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	1.85	0.998	1.12	2.54	1.85	1.66
	tak	1.36	1.99	1.33	1.68	2.47	0.645
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	7.00	5.00	7.00	10.0	8.00	7.00
	tak	7.00	8.00	5.00	6.00	9.00	4.00
SKE	nie	1.02	3.46	4.48	1.40	1.38	1.39
	tak	3.18	2.92	2.49	1.35	1.11	3.84
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.158	11.3	21.9	1.20	1.40	1.31
	tak	11.1	7.50	4.56	0.578	0.0122	15.8
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.789	0.349	0.348	0.767	0.795	0.756
	tak	0.488	0.381	0.384	0.688	0.790	0.357
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 30 Statystyki opisowe dla piątego pytania „Jaki procent swoich oszczędności byłbyś skłonny zainwestować na polecanej platformie po obejrzeniu tego nagrania?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak we wcześniejszej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości skośności, w każdym przypadku większe od 1 oraz większe od wartości błędu standardowego tej miary. Ponadto świadczy o tym wartości bezwzględne kurtozy, gdzie poza filmem 4 i 5 (grup, które twierdziły, iż rozpoznały deepfake) wartości te są większe niż wartości z błędu standardowego kurtozy. Świadczą o tym również niskie wyniki testu Shapiro-Wilk (we wszystkich przypadkach $p < 0,001$).

Histogramy poniżej obrazują rozkład odpowiedzi dla każdego filmu, z podziałem na grupy zgodnie ze zmienną nominalną E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").



Wykres 33 Zbiór 6 wykresów odpowiedzi na piąte pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazuje różnice w odpowiedzi na piąte pytanie, między osobami, które uważają, iż udało im się rozpoznać

deepfake, a tymi, które uważają, że tego nie zrobiły. Różnicę tę widać zwłaszcza dla filmów 1 i 6 (deepfake), gdzie osoby, które nie rozpoznały deepfake były skore przekazać większy procent swoich oszczędności do fałszywej inwestycji. Zauważyć można zdecydowaną przewagę wyżej punktowanych odpowiedzi, dla grupy która nie rozpoznała, iż są to nagrania fałszywe. Zastanawiająca może być korelacja wyników odpowiedzi dla filmu pierwszego z nagraniami średnio znanych influencerów, gdyż ich układ zdaje się być podobny do odpowiedzi z czwartego i piątego nagrania.

Aby określić, czy różnice są istotne statystycznie, przeprowadzono serię testów Kruskala-Wallisa. Wyniki tych testów zostały przedstawione w poniższej tabeli. Stwierdzono, że tylko w przypadku pierwszego i ostatniego filmu (oba deepfake) różnice w rozkładach są istotne statystycznie ($p < 0,05$), co oznacza, że różnice te są wystarczająco duże, aby mogły być uznane za rzeczywiste na podstawie statystyki.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_05	4.66127	1	0.031	0.08324
film2_05	0.00318	1	0.955	5.68e-5
film3_05	0.07095	1	0.790	0.00127
film4_05	1.60342	1	0.205	0.02863
film5_05	0.17804	1	0.673	0.00318
film6_05	8.42076	1	0.004	0.15037

Tabela 31 Test Kruskal-Wallis dla odpowiedzi do pytania piątego.

W celu weryfikacji hipotezy, która sugeruje, że odpowiedzi na pytanie piąte („jaki procent swoich oszczędności byłbyś skłonny zainwestować na polecanej platformie po obejrzeniu tego nagrania?") mogą różnić się w zależności od filmu, przeprowadzono test Friedmana. Wyniki tego testu wykazały, że istnieją istotne różnice w odpowiedziach na to pytanie między wszystkimi filmami (zarówno prawdziwymi, jak i deepfake) (χ^2 (5) = 89,4; $p < 0,001$). Aby zbadać różnice pomiędzy poszczególnymi odpowiedziami, przeprowadzono również test porównań par – Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_05	-	film2_05	3.700	< .001
film1_05	-	film3_05	4.002	< .001
film1_05	-	film4_05	2.455	0.015
film1_05	-	film5_05	4.473	< .001

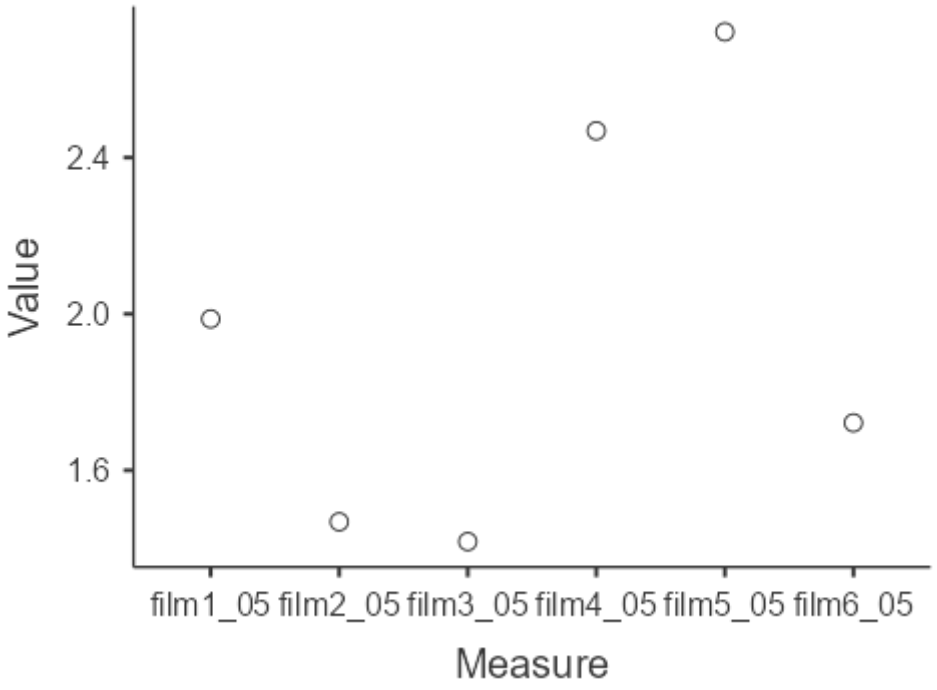
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_05	-	film6_05	1.446	0.149
film2_05	-	film3_05	0.303	0.762
film2_05	-	film4_05	6.155	< .001
film2_05	-	film5_05	8.173	< .001
film2_05	-	film6_05	2.253	0.025
film3_05	-	film4_05	6.458	< .001
film3_05	-	film5_05	8.476	< .001
film3_05	-	film6_05	2.556	0.011
film4_05	-	film5_05	2.018	0.044
film4_05	-	film6_05	3.902	< .001
film5_05	-	film6_05	5.920	< .001

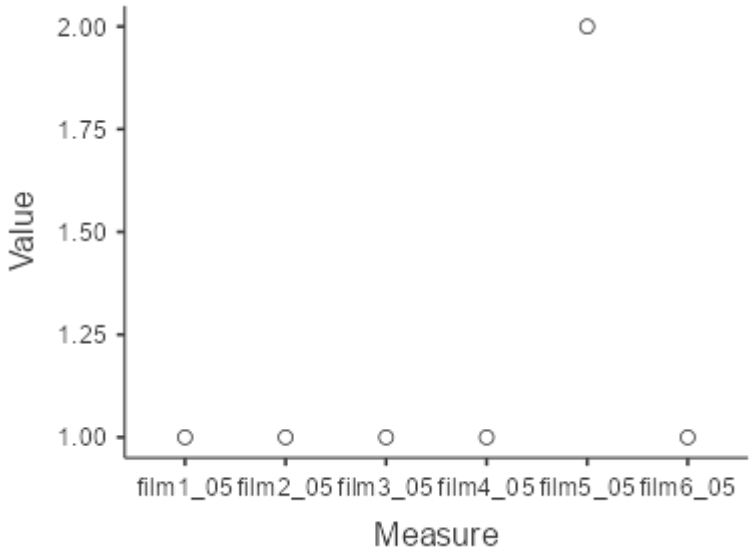
Tabela 32 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania piątego.

Powyższe obliczenia pokazały, że ilość pieniędzy, które respondenci są skorzy przekazać po obejrzeniu pierwszego nagrania różni się statystycznie od skłonności po obejrzeniu filmu 2, 3, 4 i 5, na co wskazuje $p < 0,05$. Natomiast nie zaobserwowano różnicy pomiędzy wynikami pierwszego filmu, a nagrania szóstego (deepfake) ($p > 0,05$). Pomiędzy nagraniami 2, a 3 również nie zaobserwowano różnic (oba prezentowały nieznane osoby). Zaobserwowano natomiast różnice pomiędzy filmami 2 i 3, a 4, 5 i 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również zaobserwowano różnice. Podobnie pomiędzy nimi, a nagraniem 6 (deepfake).

Ponieważ jedynie nagrania 1 i 6 oraz 2 i 3 nie były od siebie istotnie statystycznie różne, poniżej zaprezentowano wykresy średnich i median dla nagrań. Średnia prezentowana na poniższym wykresie pokazuje, iż procent oszczędności skorych do przekazania w filmach 4 i 5 były podobnie na podobnym poziomie, jednak w rzeczywistości są one od siebie istotnie różne, co dobrze pokazuje mediana. Również po medianie widać, iż wszystkie filmy poza 5 zostały zauważalnie blisko ocenione pod względem przekazania oszczędności do inwestycji.



Wykres 34 Średnia odpowiedzi dla pytania piątego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 35 Mediana odpowiedzi dla pytania piątego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Wnioskować należy, iż respondenci w zbliżony sposób oceniali swoją gotowość inwestycyjną po obejrzeniu każdego z nagrań. Jedynie film piąty (średnio znany influencer) zdaje się w niewielkim stopniu przekonywać do przekazania niewielkiej

ilości pieniędzy. Badane osoby w niewielkim stopniu są skorzy do przekazania swoich środków wyłącznie na podstawie samej reklamy wideo.

Jest to wniosek o tyle istotny, iż wskazywać może, że dopiero kolejne kroki oszustwa inwestycyjnego wpływają na gotowość danej osoby do inwestowania, a zwłaszcza kwotę jaką gotowa jest przekazać na inwestycję. Reklama ma wyłącznie skupić uwagę oglądającego i zachęcić go do zainteresowania fałszywą platformą. Na etapie rejestracji oglądający nagranie nie są świadomi, iż w krótkim okresie czasowym zostaną zmanipulowani przez oszustów do powierzenia im swoich wszystkich pieniędzy. Pozostawienie swoich danych kontaktowych na fałszywej stronie wynikać więc może bardziej z samej ciekawości i zainteresowania niż zawierzenia słowom osoby z reklamy o gwarantowanym, dużym i bezstratnym zysku.

5.3 Przekonanie co do realności zysku

Oprócz przekonania do inwestycji oraz wskazaniu procenta środków pieniężnych potencjalnie na nią przeznaczonych, zdecydowano się również odpytać respondentów o odczucie realności obiecywanego zysku. Osoby na nagraniach zapewniały bowiem o stuprocentowej skuteczności ich platformy inwestycyjnej oraz gwarantowały duże zyski bez ryzyka. Było to podkreślane zwłaszcza dzięki zestawieniu takich słów jak „inflacja" – „bezpieczny zysk" – „ciężkie czasy" – „covid-19" – „kryzys" – podkreślające wyjątkowe bezpieczeństwo platformy. Ponadto na każdym z nagrań, osoby podkreślały bezpieczeństwo swoim autorytetem, zapewniając, że same inwestują na danej platformie.

Celem pomiaru tej zmiennej, respondentom zadano pytanie „w jakim stopniu wierzysz w realność obiecywanego zysku?". Odpowiedź udzielano na dziesięciostopniowej skali, gdzie 1 to w ogóle, 10 bardzo.

Założono, iż wyniki będą zbieżne z wynikami jak dla dwóch poprzednich pytań, dotyczących inwestycji – pytania o przekonanie do inwestycji oraz wymiar potencjalnie przekazanych środków. Nagrania 4 i 5 (średnio znani influencerzy) miałyby w największym stopniu poświadczać realność zysku. Nieco niżej oceniane powinny być nagrania deepfake (1 i 6), jednak zdecydowanie lepiej niż nagrania prezentujące nieznane osoby.

Poniżej zaprezentowano tabelę z statystykami opisowymi dla szóstego pytania dla wszystkich filmów.

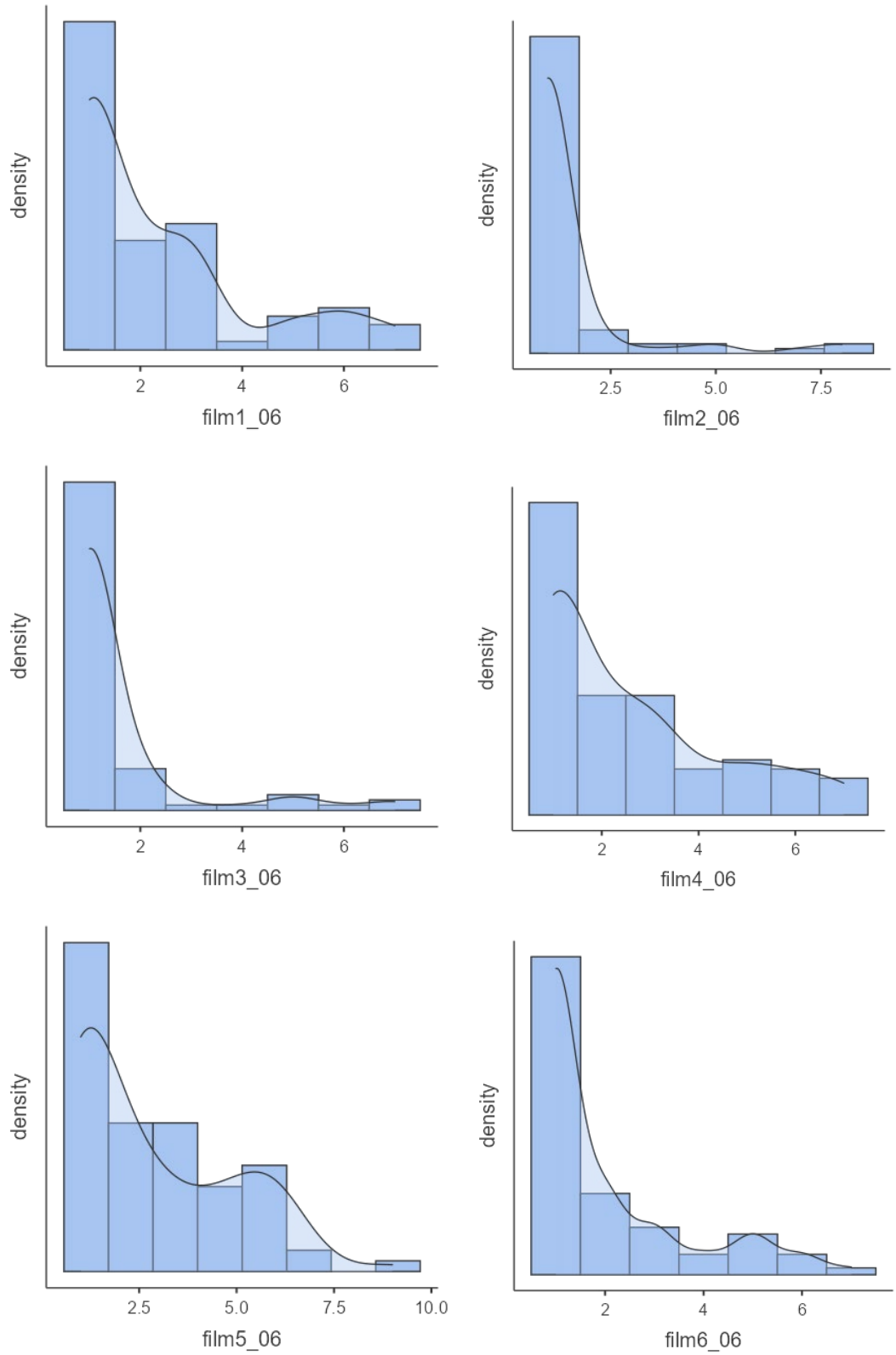
Statystyki opisowe

	film1_06	film2_06	film3_06	film4_06	film5_06	film6_06
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	2.31	1.48	1.53	2.59	2.88	2.01
SE	0.195	0.162	0.153	0.208	0.229	0.177
95% CI dolna granica przedziału ufności dla średniej	1.93	1.16	1.23	2.18	2.43	1.67
95% CI górna granica przedziału ufności dla średniej	2.70	1.79	1.83	3.00	3.32	2.36
Me	2.00	1.00	1.00	2.00	2.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.75	1.45	1.36	1.86	2.05	1.57
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	7.00	8.00	7.00	7.00	9.00	7.00
SKE	1.36	3.52	2.90	1.01	0.821	1.54
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	0.852	12.1	7.77	-0.112	-0.420	1.31
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.752	0.374	0.451	0.809	0.834	0.693
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 33 Statystyki opisowe dla szóstego pytania „W jakim stopniu wierzysz w realność obiecywanego zysku?”.

Dane z powyższej tabeli pokazują, iż nie można przyjąć, że rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p < 0{,}001$ oraz wartości skośności, które w przypadku każdego filmu są większe od wartości błędu standardowego skośności, a w większości przypadków (wyjątek stanowi wartość skośności filmu 5) są one większe od 1. Świadczą o tym również wartości kurtozy, które w większości przypadków (wyjątki stanowią filmy 4 i 5) są większe od wartości błędu standardowego tej miary. W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Na poniższych wykresach znajdują się histogramy otrzymanych wyników – dla każdego filmu osobno. Pytanie szóste brzmiało: „w jakim stopniu wierzysz w realność obiecywanego zysku”. Histogramy te również pokazują, iż rozkład wyników dla żadnego z filmów nie jest zbliżony do rozkładu normalnego.



Wykres 36 Zbiór 6 wykresów odpowiedzi na szóste pytanie dla każdego z sześciu filmów.

Na wykresach widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują silną prawoskośność dla rozkładów, z znaczną dominacją odpowiedzi 1 – wcale. Respondenci zdają się nie wierzyć w realność obiecywanego zysku. Znaczne

zróżnicowanie odpowiedzi zauważyć można jedynie dla filmów 4 i 5 (filmy prawdziwe, średnio znanych influencerów). Influencer występujący na 5 filmie zdaje się w największym stopniu przekonywać do realności wysokiego zysku.

Spośród nagrań deepfake, zwłaszcza pierwsze zdaje się przekonywać część respondentów do realności wysokich zysków. Część z respondentów w przypadku filmu 1 i 6 dokonała wysokich ocen, co sugerować może, iż nie rozpoznały one nagrania deepfake i oceniły realność zysków na poziomie zbliżonym do poziomu prezentowanego na filmach 4 i 5. Z tego powodu konieczne jest sprawdzenie czy rozpoznanie fałszu miało wpływ na ocenę realności zysków po obejrzeniu nagrań.

Tabela poniżej prezentuje dane, które są analogiczne do poprzednich, ale w tym przypadku rozkłady zmiennych zostały podzielone zgodnie ze zmienną nominalną E: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Statystyki opisowe

	zmienna nominalna E	**film1_06**	**film2_06**	**film3_06**	**film4_06**	**film5_06**	**film6_06**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	2.75	1.41	1.50	2.66	2.75	2.44
	tak	1.96	1.72	1.60	2.56	3.16	1.48
SE	nie	0.351	0.215	0.229	0.366	0.342	0.294
	tak	0.319	0.410	0.316	0.379	0.471	0.224
95% CI dolna granica przedziału ufności dla średniej	nie	2.06	0.985	1.05	1.94	2.08	1.86
	tak	1.34	0.916	0.980	1.82	2.24	1.04
95% CI górna granica przedziału ufności dla średniej	nie	3.44	1.83	1.95	3.37	3.42	3.01
	tak	2.58	2.52	2.22	3.30	4.08	1.92
Me	nie	2.00	1.00	1.00	2.00	2.00	2.00
	tak	1.00	1.00	1.00	2.00	2.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00

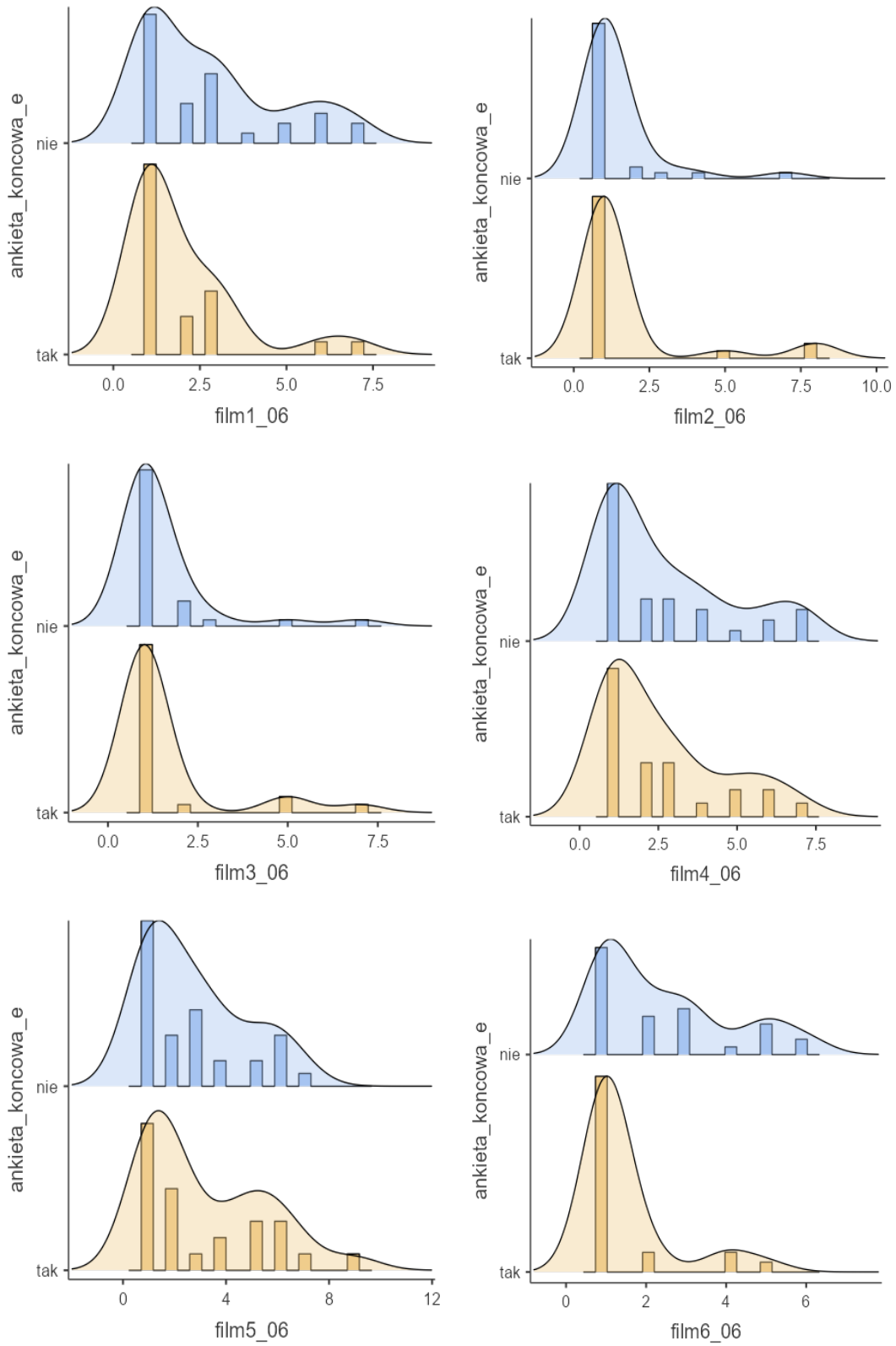
Statystyki opisowe

	zmienna nominalna E	film1_06	film2_06	film3_06	film4_06	film5_06	film6_06
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	1.98	1.21	1.30	2.07	1.93	1.66
	tak	1.59	2.05	1.58	1.89	2.36	1.12
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	7.00	7.00	7.00	7.00	7.00	6.00
	tak	7.00	8.00	7.00	7.00	9.00	5.00
SKE	nie	0.926	3.75	3.33	1.05	0.806	0.900
	tak	2.08	2.74	2.65	1.06	0.850	2.36
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.371	15.3	11.6	-0.194	-0.652	-0.451
	tak	4.30	6.35	6.18	-0.0529	-0.267	4.52
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.817	0.393	0.452	0.781	0.828	0.808
	tak	0.656	0.391	0.438	0.804	0.846	0.493
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	0.001	< .001

Tabela 34 Statystyki opisowe dla szóstego pytania „W jakim stopniu wierzysz w realność obiecywanego zysku?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości bezwzględne kurtozy, gdzie w połowie przypadków są one większe od błędu standardowego tej miary. Wyjątek stanowią filmy 4 i 5 oraz w grupie, która nie rozpoznała deepfake 1 i 6, co sugerować może, iż wyniki w tej grupie zbieżne są z wynikami oceny realności w nagraniu 4 i 5.

O braku rozkładu normalnego świadczą również wartości skośności, gdzie w każdym analizowanym przypadku wartość ta jest wyższa niż wartość z błędu standardowego tej miary. Świadczą o tym również niskie wyniki ($p \leqslant 0{,}001$) testu Shapiro-Wilk. na poniższych histogramach zilustrowano wyniki dla każdego filmu, uwzględniając podział respondentów według zmiennej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?").



Wykres 37 Zbiór 6 wykresów odpowiedzi na piąte pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Tabela oraz histogramy zaprezentowane powyżej pokazują różnice w odpowiedziach na szóste pytanie między osobami, które twierdzą, że potrafiły

rozpoznać deepfake, a tymi, które uważają, że nie dokonały tej identyfikacji. W przypadku filmów pierwszego i szóstego (oba deepfake) szczególnie widoczna jest przewaga niższych ocen w grupie, która rozpoznała, że nagranie jest fałszywe. Aby ustalić, czy te różnice są istotne statystycznie, przeprowadzono analizę za pomocą testów Kruskal-Wallis. Wyniki są przedstawione w poniższej tabeli. Stwierdzono, że tylko dla ostatniego filmu (deepfake) różnice są istotne statystycznie, co wskazuje, że rozbieżności między grupami są wystarczająco wyraźne, aby mogły być wiarygodnie analizowane za pomocą statystyk. Istotność statystyczna została wykorzystana do potwierdzenia, że różnice między grupami są na tyle duże, by nie były przypadkowe i mogły być traktowane jako prawdziwe w sensie statystycznym.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_06	2.7111	1	0.100	0.04841
film2_06	0.0454	1	0.831	8.10e-4
film3_06	0.1769	1	0.674	0.00316
film4_06	6.48e-4	1	0.980	1.16e-5
film5_06	0.2492	1	0.618	0.00445
film6_06	7.2244	1	0.007	0.12901

Tabela 35 Test Kruskal-Wallis dla odpowiedzi do pytania szóstego.

Z powyższej tabeli wnioskować można, iż respondenci deklarujący rozpoznanie fałszywych nagrań odpowiadali na to pytanie w odmienny sposób niż osoby, które fałszu nie rozpoznały wyłącznie dla nagrania szóstego. Co ciekawe, nie zaobserwowano istotnych różnic w odpowiedziach na pytanie szóste, dla nagrania pierwszego ($p > 0,05$). W tym przypadku nie należy więc mówić, o różnicy pomiędzy grupami i różnicy w wpływie nagrania na wiarę w realność obiecywanego zysku.

W celu weryfikacji hipotezy dotyczącej różnic w odpowiedziach na pytanie szóste („w jakim stopniu wierzysz w realność obiecywanego zysku?") przeprowadzono test Friedmana. Wyniki wykazały istotne różnice w odpowiedziach między wszystkimi filmami (zarówno autentycznymi, jak i deepfake), z wartością χ^2 (5) = 91,7; $p < 0,001$.

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na pytanie szóste, przeprowadzono testy porównania parami – Durbin-Conover. Tabelę z wynikami testu zamieszczono poniżej.

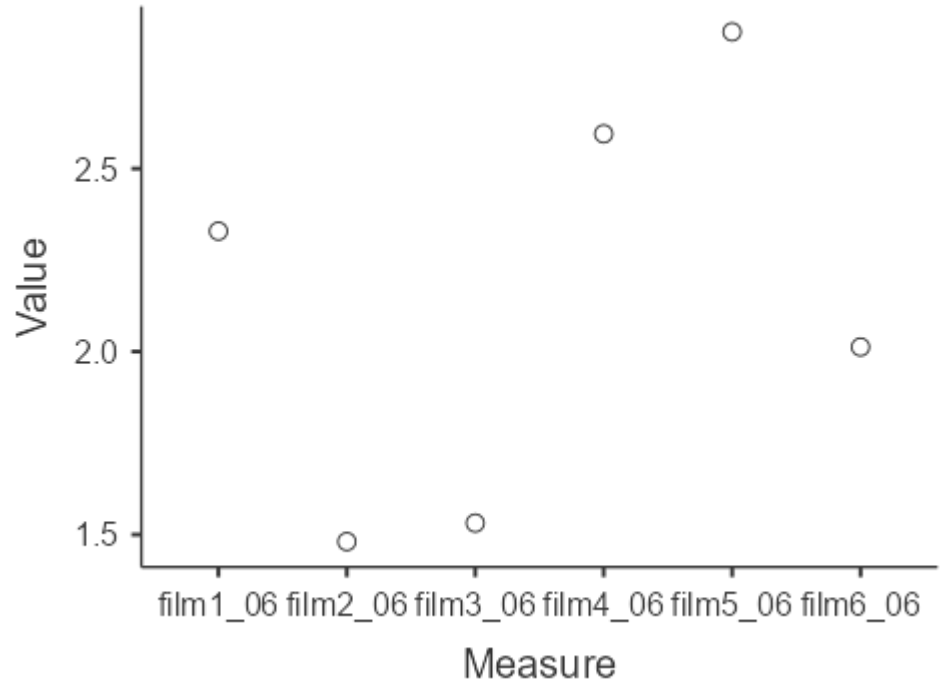
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_06	-	film2_06	5.178	< .001
film1_06	-	film3_06	4.961	< .001
film1_06	-	film4_06	1.550	0.122
film1_06	-	film5_06	3.163	0.002
film1_06	-	film6_06	2.201	0.028
film2_06	-	film3_06	0.217	0.828
film2_06	-	film4_06	6.728	< .001
film2_06	-	film5_06	8.341	< .001
film2_06	-	film6_06	2.977	0.003
film3_06	-	film4_06	6.511	< .001
film3_06	-	film5_06	8.124	< .001
film3_06	-	film6_06	2.760	0.006
film4_06	-	film5_06	1.612	0.108
film4_06	-	film6_06	3.752	< .001
film5_06	-	film6_06	5.364	< .001

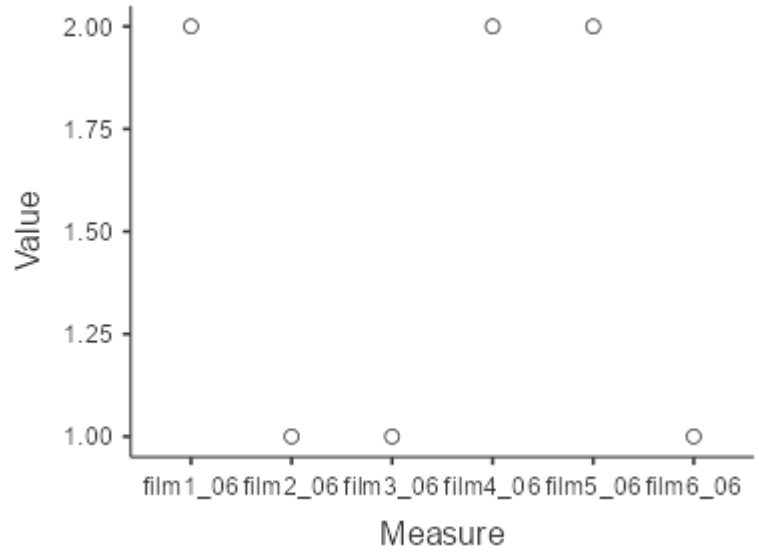
Tabela 36 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania szóstego.

Powyższe obliczenia pokazały, że film 1 różni się pod względem odbioru wiary w realność obiecywanego zysku, od filmu 2, 3, 5 i 6, na co wskazuje $p < 0,05$, natomiast nie zaobserwowano różnicy pomiędzy nim, a filmem 4 ($p > 0,05$). Potwierdza to wcześniejsze założenia, iż film wykonany z użyciem technologii deepfake, prezentujący znaną osobę, może być podobnie odbierany przez osoby nieświadome oszustwa jak pozostałe nagrania (prawdziwe) również wykorzystujące wizerunki znanych osób. Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4, 5 oraz 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również nie zaobserwowano różnicy. Film 4 i 5 są natomiast istotnie statystyczne różne od filmu 6.

Wnioskować można, iż film 1 nie był istotnie różny od filmu 4, co może świadczyć o ich podobnym odbiorze przez widzów. Aby zweryfikować tę hipotezę, poniżej zamieszczono wykresy przedstawiające zarówno średnie, jak i mediany. Jest to spowodowane tym, że rozkłady wyników znacząco odbiegają od rozkładu normalnego. W związku z tym średnie mogą być niewłaściwie reprezentatywną miarą tendencji centralnej, dlatego zaleca się stosowanie mediany.



Wykres 38 Średnia odpowiedzi dla pytania szóstego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 39 Mediana odpowiedzi dla pytania szóstego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Średnia pokazuje zbliżone wyniki filmów 2 i 3 oraz znacznie wyższe nagrań 1, 4 i 5. Porównując ich mediany, okazuje się, iż filmy 2 oraz 3 i 6 nie wzbudziły wiary co do realności obiecywanego zysku. Natomiast przekaz z nagrania 1 (deepfake) oraz 4 i 5 są zauważalnie blisko oceniane, co widać zarówno po średniej jak i medianie (Me = 2).

Wiedząc, iż film 1 nie był istotnie różny od filmu 4, zakładać można ich zbliżony efekt odbioru w aspekcie wiary w realność obiecywanego zysku. Porównując średnie punktacje grupy, która oświadczyła, iż nie rozpoznała nagrań deepfake, okazuje się, iż dla filmu pierwszego był to wynik taki sam jak dla filmu piątego, a niewiele większy od średniej dla nagrania czwartego (M1 = 2,75, M4 = 2,66, M5 = 2,75). Odczucie w aspekcie wiary w obiecywany zysk było więc niezwykle wysokie, nawet na tle prawdziwych nagrań średnio znanych influencerów.

5.4 Obawa o utratę pieniędzy

Odpowiedzialny inwestor, chcący przeznaczyć część swoich oszczędności celem jego pomnażana, zapoznając się z informacjami dotyczącymi inwestycji, przygląda się zwłaszcza ryzyku inwestycyjnemu. W niniejszym badaniu, na każdym z nagrań, osoby na nich prezentowane gwarantowały całkowite bezpieczeństwo ich giełdy, a także sugerowały brak ryzyka dla umieszczonych tam pieniędzy. Potwierdzały to własnym autorytetem, twierdząc, iż same z sukcesem inwestują na tych platformach.

Postanowiono zbadać jak poszczególne osoby, z różnym doświadczeniem życiowym, podchodzą do ryzyka inwestycyjnego. Czy gwarancja znanego lub nieznanego influencera może być uznawana za wystarczającą? do pomiaru tej zmiennej zdecydowano się zadać respondentom pytanie „po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?”. Odpowiedzi, podobnie jak w przypadku poprzednich pytań, udzielone zostały na 10 stopniowej skali. Odpowiedź jeden to „w ogóle” odpowiedź 10 to „bardzo”.

Założono, iż w przypadku prawdziwych filmów średnio znanych influencerów obawa o utratę środków będzie najmniejsza, choć nadal wysoka. Podobnie odbiór ryzyka określony powinien być wśród filmów deepfake, gdzie zakłada się, iż na średnią obawy znacząco może wpłynąć grupa osób, która rozpoznała fałsz nagrań. Oczywistym zdaje się być fakt, iż w przypadku rozpoznania deepfake, oglądający przestraszy się, iż może paść ofiarą oszustwa i oceni ryzyko utraty środków na najwyższym poziomie. Założono, iż filmami, po których respondenci będą obawiać się utraty swoich pieniędzy, będą nagrania prezentujące nieznane osoby. Nieznane twarze, słabej jakości gra aktorska oraz niskiej jakości wideo, mogą skutecznie wpłynąć na odbiór braku profesjonalizmu „inwestora”, a za tym na wyczucie oszustwa przez respondentów.

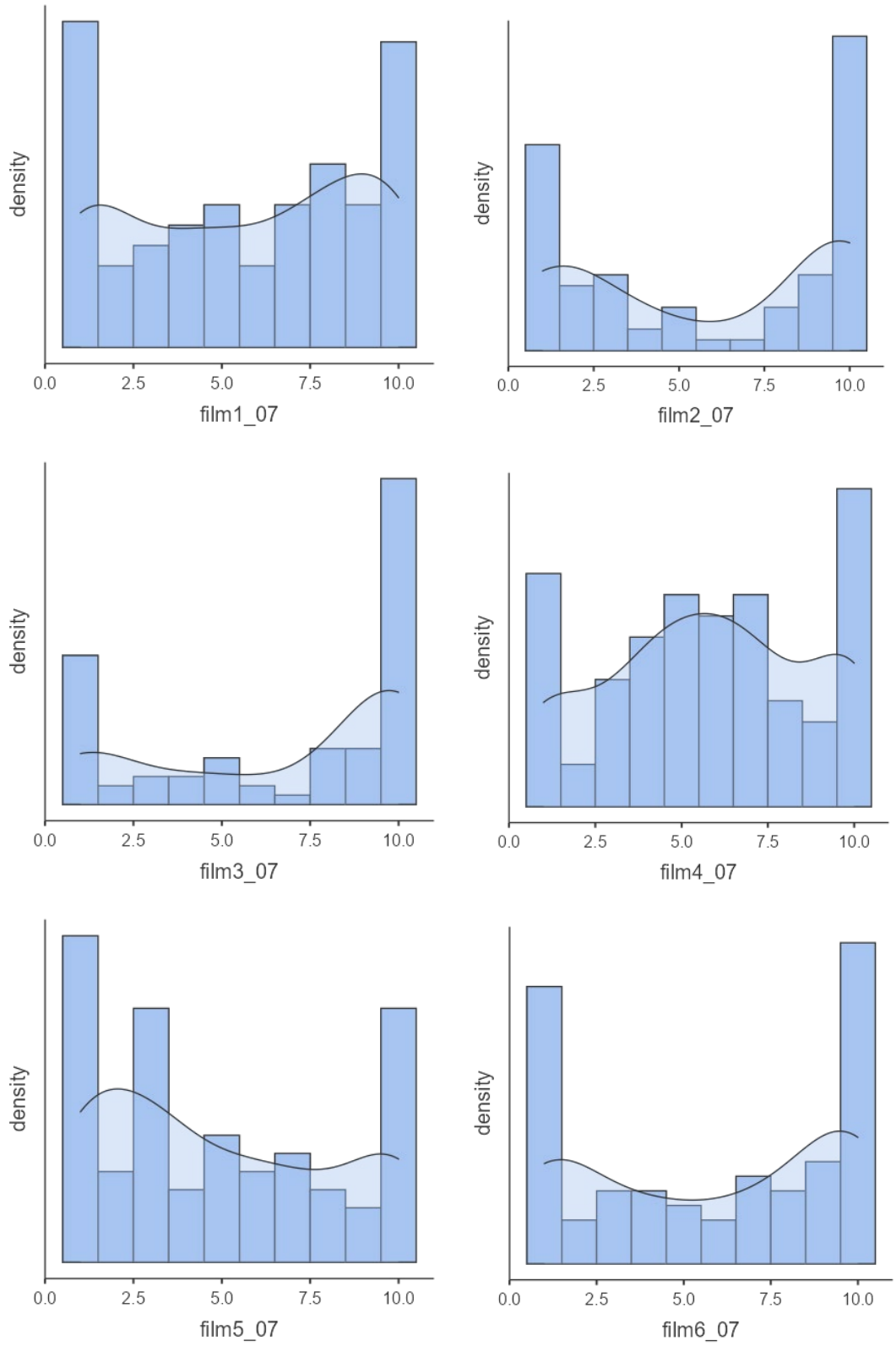
W poniższej tabeli zaprezentowano statystyki opisowe dla odpowiedzi na pytanie siódme dla każdego z nagrań.

Statystyki opisowe

	film1_07	**film2_07**	**film3_07**	**film4_07**	**film5_07**	**film6_07**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	5.70	5.97	6.80	5.81	4.90	5.86
SE	0.373	0.434	0.416	0.334	0.368	0.409
95% CI dolna granica przedziału ufności dla średniej	4.97	5.12	5.98	5.16	4.18	5.06
95% CI górna granica przedziału ufności dla średniej	6.43	6.83	7.61	6.47	5.62	6.66
Me	6.00	7.50	9.00	6.00	4.00	7.00
D	1.00	10.0	10.0	10.0	1.00	10.0
SD	3.34	3.88	3.69	2.99	3.29	3.64
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	10.0	10.0	10.0	10.0	10.0
SKE	-0.149	-0.178	-0.619	-0.108	0.349	-0.184
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-1.45	-1.79	-1.35	-1.08	-1.31	-1.62
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.884	0.781	0.761	0.923	0.877	0.835
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 37 Statystyki opisowe dla siódmego pytania „Po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?”.

Powyższa tabela przedstawia statystyki opisowe odpowiedzi na siódme pytanie: „Po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?”. Z analizy danych wynika, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Wartości testu Shapiro-Wilk, które przyjmują $p < 0{,}001$ oraz bezwzględne wartości kurtozy, które dla każdego filmu przekraczają błąd standardowy tej miary, a w większości przypadków są większe niż 1, potwierdzają tę tezę. W związku z brakiem rozkładu normalnego uzyskanych wyników, w dalszych analizach zastosowano testy nieparametryczne.



Wykres 40 Zbiór 6 wykresów odpowiedzi na siódme pytanie dla każdego z sześciu filmów.

Na wykresach 1, 2, 3, 5 i 6 widzimy symetrycznie przedstawione skrajne odpowiedzi. Uzyskane dane wskazują w tych przypadkach na układ dwumodalny, zwany również bimodalnym, z znaczną dominacją odpowiedzi 1 – wcale oraz 10 – bardzo.

Dzieje się tak zwłaszcza dla filmu 1, 5 i 6 gdzie film 1 i 6 były filmami deepfake, zaś film 5 był prawdziwy i prezentował średnio znanego influencera. Wykres 4, również prezentujący średnio znanego influencera, ma najbardziej zrównane wyniki odpowiedzi i w swojej budowie wskazuje na rozkład płaski, zwany również rozkładem równomiernym.

Podkreślić należy, iż rozkład dwumodalny jest to rozkład, w przypadku którego występują dwie wartości modalne. Specyficzne ułożenie maksim, względem rozkładów odpowiedzi na pozostałe pytania, stanowić może ważne informacje na temat istoty badanej zmiennej. Dwumodalność może w tym przypadku wskazywać na występowanie rozbieżności opinii, niejednorodność próby lub problemy z narzędziem pomiarowym. Może również świadczyć o występowaniu obciążonych odpowiedzi.

W celu dalszej weryfikacji odpowiedzi i ustalenia rzeczywistego stanu, zdecydowano się na kontynuację analizy, uwzględniając zmienną nominalną E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Poniższa tabela prezentuje dane analogiczne do tych z wcześniejszej tabeli, przy czym rozkłady zmiennych zostały podzielone według zmiennej nominalnej E.

Statystyki opisowe

	zmienna nominalna E	**film1_07**	**film2_07**	**film3_07**	**film4_07**	**film5_07**	**film6_07**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	5.59	7.09	7.03	5.50	5.31	5.09
	tak	6.24	5.88	6.04	5.80	5.80	5.80
SE	nie	0.583	0.665	0.631	0.537	0.569	0.636
	tak	0.705	0.799	0.820	0.643	0.658	0.798
95% CI dolna granica przedziału ufności dla średniej	nie	4.45	5.79	5.79	4.45	4.20	3.85
	tak	4.86	4.31	4.43	4.54	4.51	4.24

Statystyki opisowe

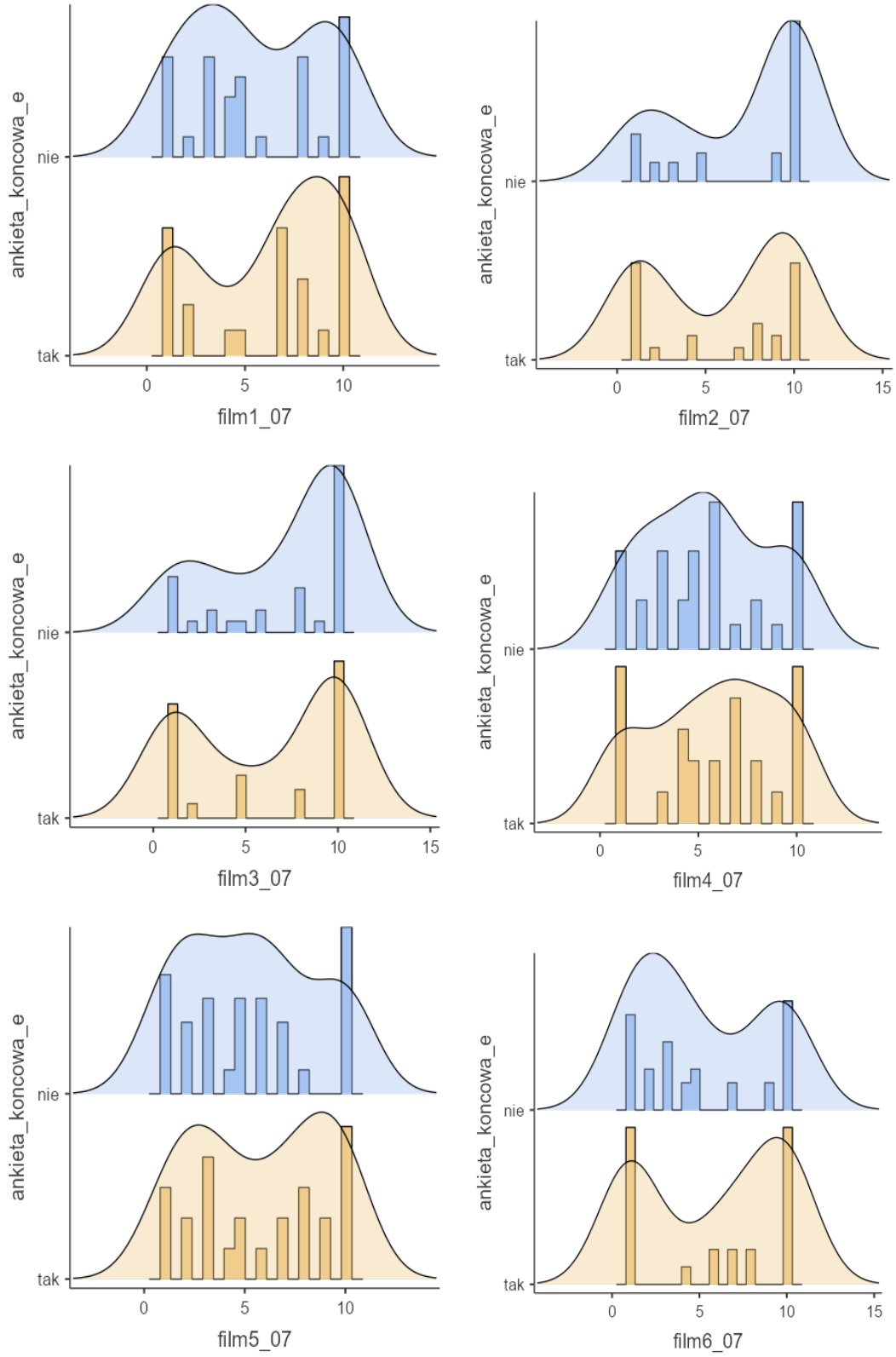
	zmienna nominalna E	film1_07	film2_07	film3_07	film4_07	film5_07	film6_07
95% CI górna granica przedziału ufności dla średniej	nie	6.74	8.40	8.27	6.55	6.43	6.34
	tak	7.62	7.45	7.65	7.06	7.09	7.36
Me	nie	5.00	10.0	8.50	5.50	5.00	4.00
	tak	7.00	8.00	8.00	6.00	6.00	7.00
D	nie	10.0	10.0	10.0	6.00 [a]	10.0	10.0
	tak	10.0	1.00 [a]	10.0	1.00 [a]	10.0	1.00 [a]
SD	nie	3.30	3.76	3.57	3.04	3.22	3.60
	tak	3.53	3.99	4.10	3.21	3.29	3.99
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	10.0	10.0	10.0	10.0	10.0
	tak	10.0	10.0	10.0	10.0	10.0	10.0
SKE	nie	0.0734	-0.725	-0.749	0.129	0.201	0.341
	tak	-0.489	-0.250	-0.260	-0.221	-0.0823	-0.221
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-1.47	-1.34	-1.13	-1.11	-1.24	-1.56
	tak	-1.36	-1.87	-1.86	-1.18	-1.54	-1.80
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.886	0.714	0.767	0.920	0.899	0.835
	tak	0.834	0.773	0.748	0.904	0.894	0.776
PS-W	nie	0.003	< .001	< .001	0.021	0.006	< .001
	tak	< .001	< .001	< .001	0.022	0.013	< .001

[a] Istnieje więcej niż jedno D, tylko pierwsze jest odnotowywane.

Tabela 38 Statystyki opisowe dla siódmego pytania „Po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości testu Shapiro-Wilk, we wszystkich filmach $p < 0,05$ oraz wartości bezwzględne kurtozy, w każdym przypadku większe od 1 oraz większe od błędu standardowego tej miary. Poniżej znajdują się histogramy dla każdego z filmów. Podobnie jak w powyższej tabeli, respondenci zostali podzieleni na podstawie

zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.



Wykres 41 Zbiór 6 wykresów odpowiedzi na siódme pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na siódme pytanie, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmów 1 i 6 (deepfake) oraz 2 i 3 (nieznane osoby) zauważyć można lekką przewagę wyżej ocenianych odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe.

Aby ocenić istotność statystyczną różnic między grupami, zrealizowano serię testów Kruskal-Wallis. Rezultaty tych testów zostały przedstawione w poniższej tabeli.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_07	0.334	1	0.563	0.00597
film2_07	2.478	1	0.115	0.04424
film3_07	0.682	1	0.409	0.01218
film4_07	0.171	1	0.679	0.00306
film5_07	0.287	1	0.592	0.00512
film6_07	0.177	1	0.674	0.00316

Tabela 39 Test Kruskal-Wallis dla odpowiedzi do pytania siódmego.

Na podstawie powyższej tabeli wnioskować można, iż nie zaobserwowano istotnych statystycznie różnic pomiędzy grupami. W żadnym przypadku nie stwierdzono tym samym różnic pomiędzy dwoma grupami na tyle dużych, aby móc je wiarygodnie wyjaśnić przy użyciu statystyk.

Aby zweryfikować hipotezę dotyczącą różnic w odpowiedziach na pytanie siódme („po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?") w kontekście różnych filmów, przeprowadzono test Friedmana, będący nieparametrycznym odpowiednikiem analizy wariancji z powtarzalnymi pomiarami (RM ANOVA). Wyniki wskazują, że odpowiedzi na to pytanie różnią się pomiędzy wszystkimi filmami (zarówno rzeczywistymi, jak i fałszywymi) (χ^2 (5) = 26,2; p < 0,001). W celu analizy różnic pomiędzy poszczególnymi odpowiedziami na pytanie siódme, zastosowano test porównania parami – Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_07	-	film2_07	0.798	0.425
film1_07	-	film3_07	2.985	0.003

Porównania Parami (Durbin-Conover)

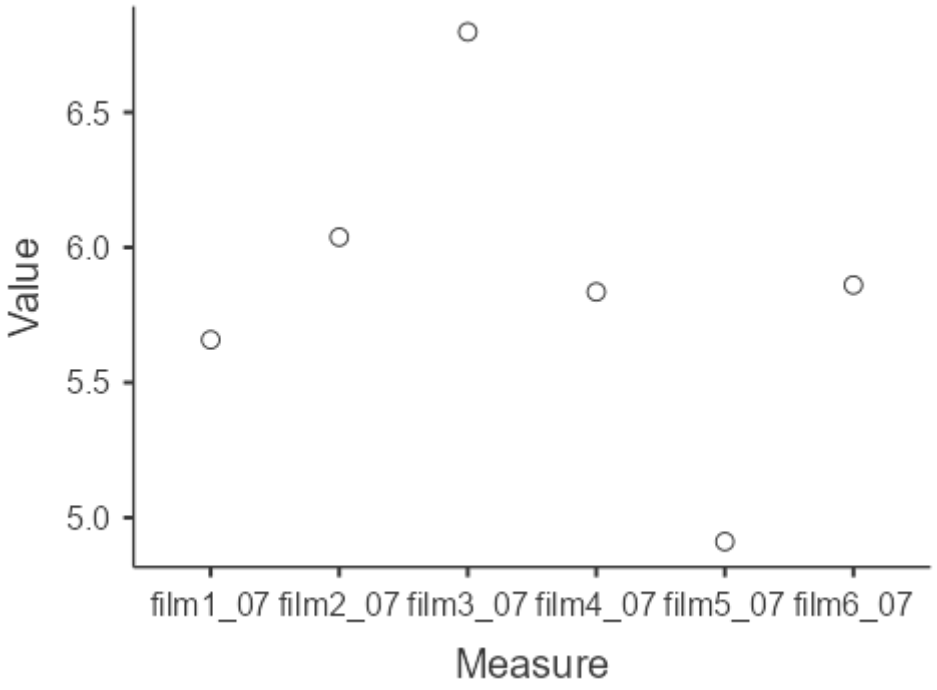
			Statistic	p
film1_07	-	film4_07	0.412	0.681
film1_07	-	film5_07	2.110	0.035
film1_07	-	film6_07	0.283	0.777
film2_07	-	film3_07	2.187	0.029
film2_07	-	film4_07	1.210	0.227
film2_07	-	film5_07	2.908	0.004
film2_07	-	film6_07	0.515	0.607
film3_07	-	film4_07	3.397	< .001
film3_07	-	film5_07	5.095	< .001
film3_07	-	film6_07	2.702	0.007
film4_07	-	film5_07	1.698	0.090
film4_07	-	film6_07	0.695	0.488
film5_07	-	film6_07	2.393	0.017

Tabela 40 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania siódmego.

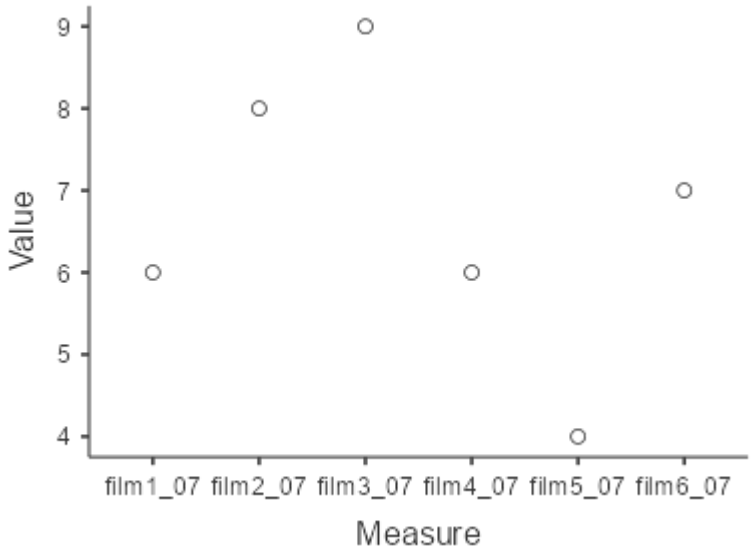
Powyższe obliczenia pokazały, że film 1 różni się od filmu 3 i 5, na co wskazuje $p < 0,05$, natomiast nie zaobserwowano różnicy pomiędzy nim, a filmem 2, 4 i 6 ($p > 0,05$). Różnice zaobserwowano również pomiędzy nagraniami 2, a 3 i 5. Nie zaobserwowano różnic pomiędzy filmem 2, a 4 i 6. Różnica jest natomiast między filmem 3, a 4, 5 i 6. Film 4 nie jest statystycznie istotnie różny w aspekcie obawy o utratę środków od filmu 5 i 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 5, a 6 (deepfake) zaobserwowano istotne statystycznie różnice.

Wnioskować można, iż film 1 nie był istotnie różny od filmów 4 i 6, co dziwne nie był jednak różny od filmu 2 (nieznana osoba), który to okazał się istotnie statystycznie różny od filmu 3 (również nieznana osoba). Zaobserwowano również istotne statystycznie różnice pomiędzy nagraniem 3, a 4 i 5 (średnio znani influencerzy).

Celem weryfikacji tych hipotez, poniżej umieszczono wykresy przedstawiające średnie oraz mediany. Wynika to z tego, iż rozkłady wyników są znacząco różne od rozkładu normalnego (rozkład płaski). Średnie są więc w tym przypadku mocno obciążoną miarą tendencji centralnej, w związku z czym zaleca się stosowanie mediany.



Wykres 42 Średnia odpowiedzi dla pytania siódmego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 43 Mediana odpowiedzi dla pytania siódmego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Jak wykazano w niniejszym podrozdziale, rozkład dwumodalny sugerować może kilka aspektów niezwykle istotnych dla badacza. Na przykład, jeśli dana zmienna reprezentuje deklarowane preferencje lub postawy, wówczas dwumodalność może wskazywać na występowanie rozbieżności opinii. W tym przypadku zdaje się to być

jednak problem z narzędziem pomiarowym oraz tendencyjnością odpowiedzi. Pytanie siódme było bowiem jedynym pytaniem, gdzie punktacja przyznawana była odwrotnie w sensie negatywnego odbioru przekazu filmowego. Na pytanie „po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?", odpowiedź 1 oznaczała „w ogóle", zaś 10 „bardzo". Część respondentów mogła odpowiedzieć tendencyjnie, zaznaczając 1 jako „bardzo się obawiam". Świadczyć ponadto o tym może ilość skrajnych odpowiedzi (1 i 10) oraz brak istotnie statystycznych różnic wewnątrzgrupowych dla nagrań deepfake, które w analizach odpowiedzi na pozostałe pytania zawsze występują. Z tego powodu nie udało się również wykazać czy wykrycie fałszu powoduje przyznanie wyższej oceny ryzyka utraty pieniędzy. Przykładem takiej pomyłki, może być treść zamieszczona w jednym z pól otwartej odpowiedzi: „w poprzednich pytaniach, błędnie zaznaczano odpowiedz „o obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?". w poprzendich miało być 10– bardzo się obawiasz".

Pomimo tego, wykazano, iż zachowana została założona tendencja odpowiedzi. Nagrania prezentujące nieznane osoby, otrzymały najniższe wyniki, zarówno pod względem średniej jak i mediany. Zarówno mediana jak i średnia wskazują jednoznacznie na wysokie oceny filmu 3 oraz niskie 5 (średnio znany influencer). Odpowiedzi do filmów 1, 2, 4 i 6 są na obu wykresach do siebie zbliżone. Brak istotnych statystycznie różnic pomiędzy tymi nagraniami potwierdziły obliczenia Durbin-Conover. Wnioskować można, iż respondenci w zbliżonym stopniu obawiają się utraty zainwestowanych środków, zarówno w przypadku nagrań deepfake jak i filmów prezentujących prawdziwych influencerów.

5.5 Wpływ nagrań na decyzje inwestycyjne innych osób

Podejmując decyzje inwestycyjne, inwestorzy często przeceniają swoją wiedzę i swoje umiejętności. Złudzenie ponadprzeciętności jest błędem poznawczym, objawiającym się skłonnością jednostki do przeceniania swoich umiejętności i cech w stosunku do innych ludzi[265]. Objawia się to zwłaszcza w momencie podejmowania ważnych dla siebie decyzji. Jedną z nich może być decyzja o inwestowaniu, przy której

[265] V. Hoorens, „Self-enhancement and Superiority Biases in Social Comparison", European Review of Social Psychology, 4 (1), 1993, s. 113–139.

jednostka może przeceniać swoje umiejętności intelektualne, a decyzje innych osób uważać za gorsze.

Zastanawiające jest, czy osoby, które same nie dały się przekonać do inwestycji, uważają, że inne osoby są w stanie się na to nabrać. Ponadto interesujące jest, jak ocenią wiedzę innych osób, respondenci, którzy rozpoznali deepfake, a jak ci którzy tego nie zrobili. Czy będą uważać, iż inni mogą się nabrać na przygotowane nagrania?

By odpowiedzieć na te pytania, postanowiono zadać respondentom następujące pytanie: „w jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycje innych osób oglądających go?". Odpowiedź udzielano na dziesięciostopniowej skali, gdzie 1 to w ogóle, 10 bardzo.

Założono, iż najwyższe wyniki otrzymają prawdziwe filmu średnio znanych influencerów. Nagrania deepfake dla osób, które ich nie rozpoznały będą równie wysoko punktowane, natomiast osoby, które je rozpoznają ocenią je jako średnio wpływające, z możliwością wzięcia pod uwagę faktu, iż inna osoba może tej różnicy nie zauważyć. Za najmniej wpływowe uznane zostaną nagrania prezentujące nieznane osoby. W poniższej tabeli zaprezentowano statystyki opisowe udzielonych odpowiedzi.

Statystyki opisowe

	film1_08	**film2_08**	**film3_08**	**film4_08**	**film5_08**	**film6_08**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	4.33	2.66	3.10	4.84	4.58	3.20
SE	0.307	0.239	0.270	0.272	0.267	0.261
95% CI dolna granica przedziału ufności dla średniej	3.72	2.19	2.57	4.30	4.05	2.69
95% CI górna granica przedziału ufności dla średniej	4.93	3.13	3.63	5.37	5.10	3.71
Me	4.50	2.00	2.00	5.00	5.00	3.00
D	1.00	1.00	1.00	1.00 [a]	5.00	1.00
SD	2.75	2.13	2.40	2.44	2.39	2.32
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	10.0	10.0	10.0	10.0	10.0
SKE	0.243	1.58	1.13	-0.210	-0.101	0.807
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-1.06	2.04	0.573	-0.710	-0.891	-0.202
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.908	0.774	0.829	0.926	0.932	0.858
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

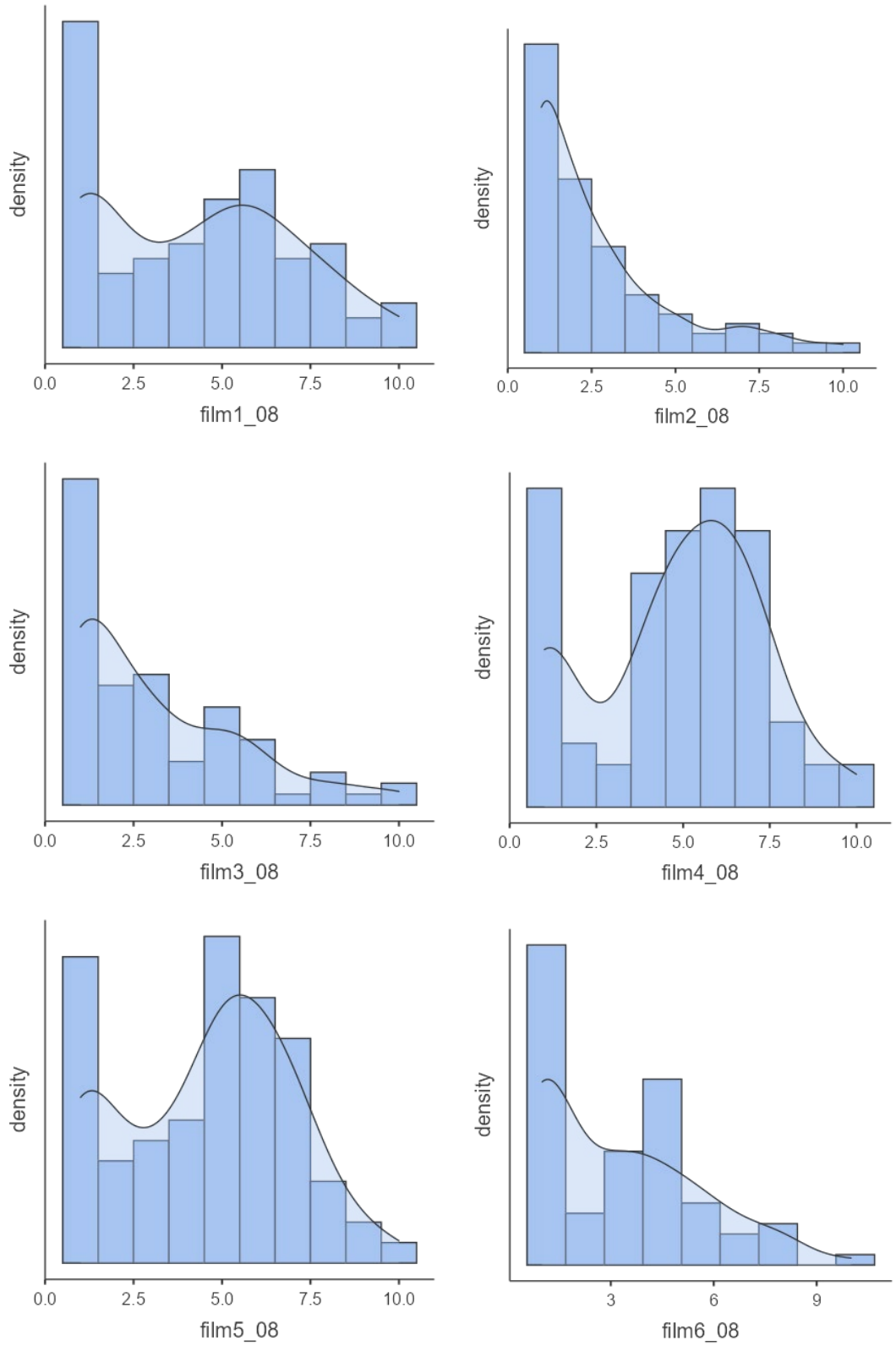
[a] Istnieje więcej niż jedno D, tylko pierwsze jest odnotowywane.

Statystyki opisowe

	film1_08	film2_08	film3_08	film4_08	film5_08	film6_08

Tabela 41 Statystyki opisowe dla ósmego pytania „W jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycje innych osób oglądających go?”.

Dane z powyższej tabeli wskazują, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Wartości testu Shapiro-Wilk potwierdzają to, osiągając $p < 0{,}001$ we wszystkich przypadkach, a wartości skośności, w połowie przypadków, przekraczają błąd standardowy tej miary, z dwiema wartościami większymi od 1. Dodatkowo, w pięciu przypadkach wartości bezwzględne kurtozy przewyższają błąd standardowy tej miary, z wyjątkiem filmu szóstego (deepfake). W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych analizach statystycznych zastosowano testy nieparametryczne. Na poniższych wykresach znajdują się histogramy wyników dla każdego filmu osobno. Pytanie ósme brzmiało: „W jakim stopniu film ten może wpłynąć na decyzje inwestycyjne innych osób oglądających go?”. Histogramy prezentujące rozkład wyników również ilustrują, że dla żadnego z filmów nie jest on zbliżony do rozkładu normalnego.



Wykres 44 Zbiór 6 wykresów odpowiedzi na ósme pytanie dla każdego z sześciu filmów.

Na wykresach 1, 2, 3 i 6 widzimy przewagę niskich odpowiedzi. Uzyskane dane wskazują w tych przypadkach silną prawoskośność, z znaczną dominacją odpowiedzi 1 – wcale. Dzieje się tak zwłaszcza dla filmu 2 i 3 gdzie oba filmy były prawdziwe i prezentowały nieznane osoby. Po wykresie nagrania pierwszego (deepfake) wyróżnić można dążenie do bimodalności (zwiększona liczba środkowych odpowiedzi). Jednak podręcznikową dwumianowość osiągnęły dopiero nagrania średnio znanych influencerów, czyli 4 i 5.

W celu analizy rozkładu wyników w kontekście rozpoznania lub braku rozpoznania nagrań deepfake, rozkłady zmiennych zostały podzielone na podstawie zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Poniższa tabela przedstawia dane w sposób analogiczny do wcześniejszej, z uwzględnieniem podziału według zmiennej nominalnej E.

Statystyki opisowe

	zmienna nominalna E	**film1_08**	**film2_08**	**film3_08**	**film4_08**	**film5_08**	**film6_08**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	5.47	2.69	2.88	5.22	5.06	3.66
	tak	3.20	2.56	2.96	3.92	4.52	2.44
SE	nie	0.484	0.363	0.361	0.403	0.356	0.398
	tak	0.500	0.504	0.546	0.568	0.542	0.469
95% CI dolna granica przedziału ufności dla średniej	nie	4.52	1.98	2.17	4.43	4.36	2.88
	tak	2.22	1.57	1.89	2.81	3.46	1.52
95% CI górna granica przedziału ufności dla średniej	nie	6.42	3.40	3.58	6.01	5.76	4.44
	tak	4.18	3.55	4.03	5.03	5.58	3.36
Me	nie	6.00	2.00	2.50	6.00	5.00	3.50
	tak	2.00	1.00	2.00	4.00	5.00	1.00
D	nie	1.00 [a]	1.00	1.00	6.00	6.00	1.00

Statystyki opisowe

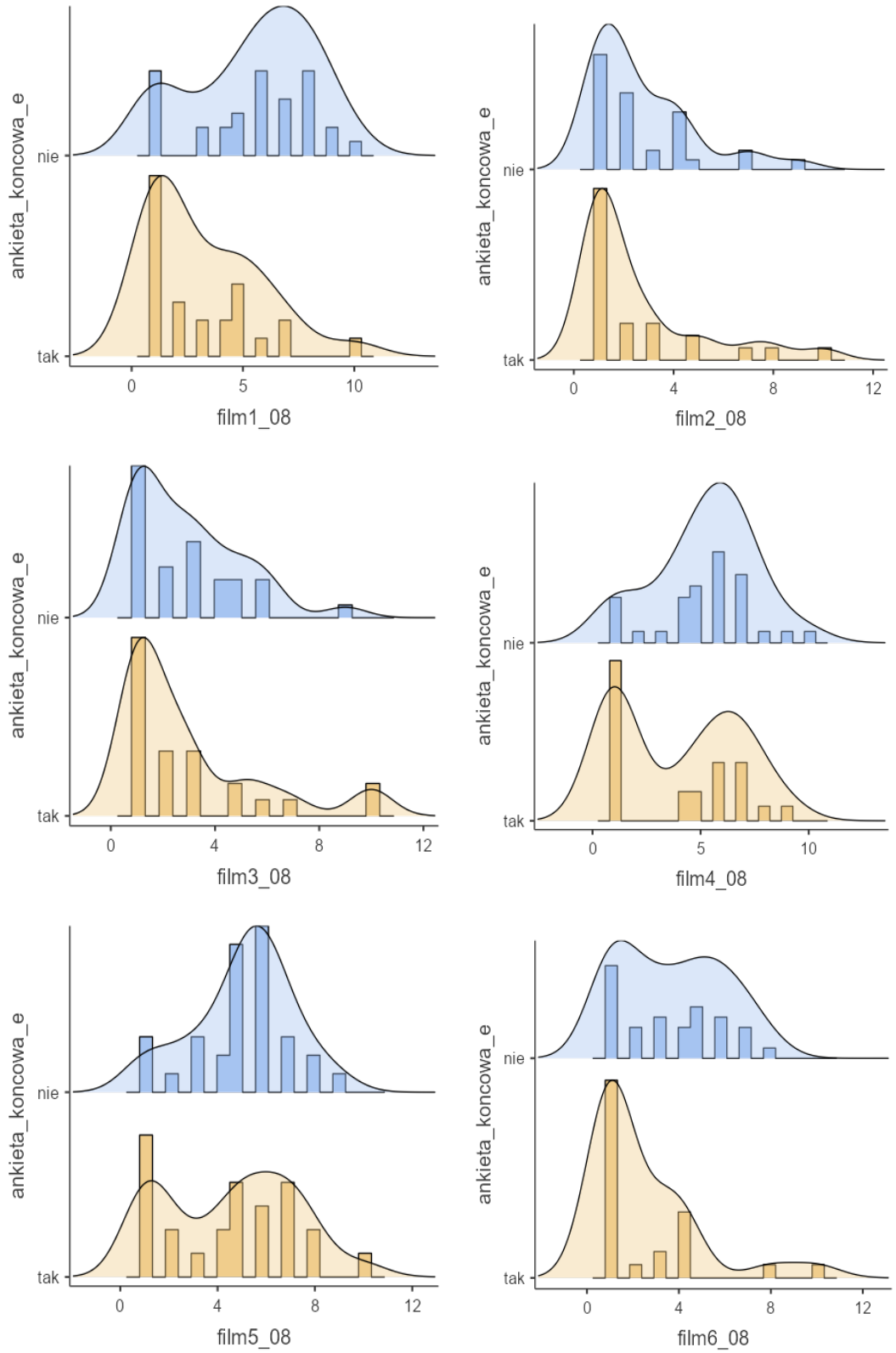
	zmienna nominalna E	film1_08	film2_08	film3_08	film4_08	film5_08	film6_08
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.74	2.05	2.04	2.28	2.02	2.25
	tak	2.50	2.52	2.73	2.84	2.71	2.35
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	9.00	9.00	10.0	9.00	8.00
	tak	10.0	10.0	10.0	9.00	10.0	10.0
SKE	nie	-0.468	1.50	1.10	-0.374	-0.495	0.224
	tak	1.04	1.80	1.61	0.159	0.0513	2.03
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.882	2.01	0.972	-0.0555	0.0120	-1.27
	tak	0.559	2.55	1.85	-1.63	-1.06	4.20
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.909	0.794	0.848	0.935	0.934	0.899
	tak	0.839	0.688	0.741	0.821	0.918	0.667
PS-W	nie	0.011	< .001	< .001	0.054	0.052	0.006
	tak	0.001	< .001	< .001	< .001	0.045	< .001

[a] Istnieje więcej niż jedno D, tylko pierwsze jest odnotowywane.

Tabela 42 Statystyki opisowe dla ósmego pytania „W jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycje innych osób oglądających go?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości kurtozy oraz skośności, gdzie w ponad połowie analizowanych przypadków wartość bezwzględna z wartości kurtozy i średniej jest wyższa niż wartość z błędu standardowego odpowiednio kurtozy bądź skośności. Świadczą o tym również niskie wyniki ($p \leqslant 0{,}05$) testu Shapiro-Wilk, z dominacją $p \leqslant 0{,}001$.

Poniższe histogramy ilustrują wyniki dla każdego z filmów. Uczestnicy badania zostali sklasyfikowani na podstawie zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".



Wykres 45 Zbiór 6 wykresów odpowiedzi na ósme pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Zarówno tabela, jak i powyższe histogramy, przedstawiają różnice w odpowiedziach na ósme pytanie między uczestnikami, którzy uważają, że rozpoznali deepfake, a tymi, którzy tego nie zauważyli. Szczególnie w przypadku pierwszego i ostatniego filmu (oba deepfake) można dostrzec wyraźną przewagę niższych ocen w grupie, która prawidłowo zidentyfikowała nagrania jako fałszywe. Osoby te uważają, że niskie jest prawdopodobieństwo, by fałszywe nagrania wpłynęły na decyzje inwestycyjne innych osób. Zastanawiająca może być korelacja wyników odpowiedzi dla filmu pierwszego z filmami 4 i 5, gdyż ich układ zdaje się być podobny do siebie.

Aby sprawdzić, czy różnice w grupie są istotne statystycznie, przeprowadzono serię testów Kruskal-Wallis. Wyniki, przedstawione w poniższej tabeli, pokazują, że istotne różnice w rozkładach występują jedynie w przypadku pierwszego i ostatniego filmu (deepfake), gdzie $p < 0,05$. Oznacza to, że różnice między tymi dwoma grupami są na tyle znaczące, że można je uznać za wiarygodne na podstawie statystyk. Istotność statystyczna wskazuje, że różnice te nie są przypadkowe i mają solidne podstawy w wynikach badania.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_08	9.064	1	0.003	0.16186
film2_08	0.873	1	0.350	0.01559
film3_08	0.168	1	0.682	0.00300
film4_08	2.524	1	0.112	0.04507
film5_08	0.514	1	0.474	0.00917
film6_08	5.748	1	0.017	0.10263

Tabela 43 Test Kruskal-Wallis dla odpowiedzi do pytania ósmego.

Na podstawie danych zawartych w tabeli można wnioskować, że respondenci, którzy rozpoznali fałszywe nagrania, odpowiadali na pytanie dotyczące filmów pierwszego i szóstego w sposób odmienny od osób, które nie rozpoznały fałszu. Statystycznie istotne różnice, zwłaszcza dla filmu szóstego, sugerują, że osoby, które nie zidentyfikowały deepfake, oceniały ten film jako mający potencjalny wpływ na decyzje inwestycyjne innych. Aby przetestować hipotezę, że odpowiedzi na pytanie ósme („w jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycyjne innych osób?") różnią się w zależności od filmu, zastosowano test Friedmana. Wyniki testu (χ^2 (5) = 76,4; $p < 0,001$) potwierdzają istotne różnice w odpowiedziach. Poniżej

zamieszczono wyniki z przeprowadzonego testu porównań parami Durbin-Conover. Wyniki przedstawiono w tabeli.

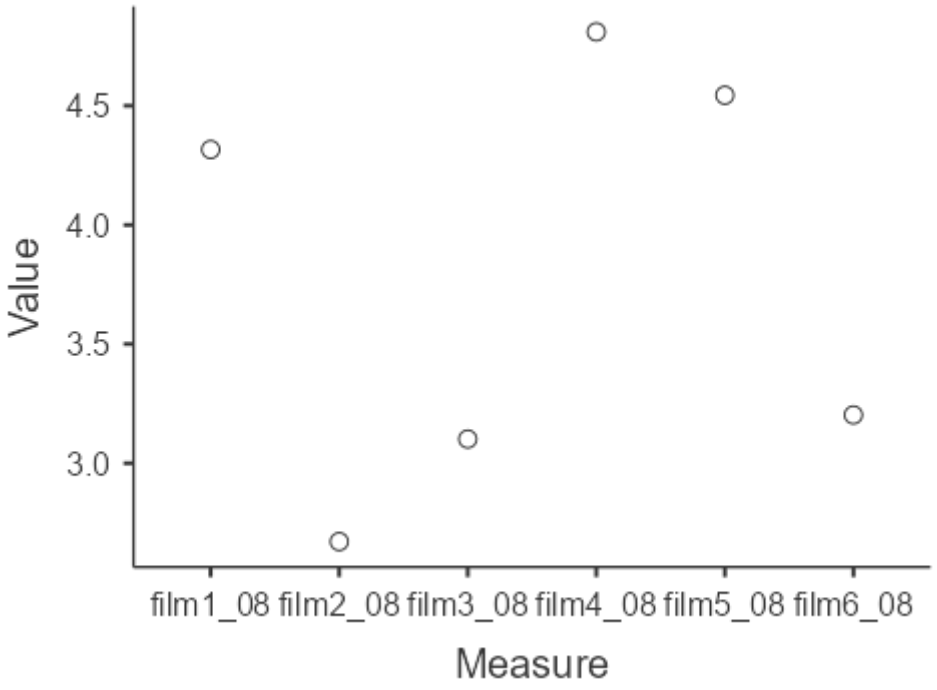
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_08	-	film2_08	6.3113	< .001
film1_08	-	film3_08	5.3079	< .001
film1_08	-	film4_08	0.3961	0.692
film1_08	-	film5_08	0.0264	0.979
film1_08	-	film6_08	4.3308	< .001
film2_08	-	film3_08	1.0035	0.316
film2_08	-	film4_08	6.7075	< .001
film2_08	-	film5_08	6.3378	< .001
film2_08	-	film6_08	1.9805	0.048
film3_08	-	film4_08	5.7040	< .001
film3_08	-	film5_08	5.3343	< .001
film3_08	-	film6_08	0.9771	0.329
film4_08	-	film5_08	0.3697	0.712
film4_08	-	film6_08	4.7269	< .001
film5_08	-	film6_08	4.3572	< .001

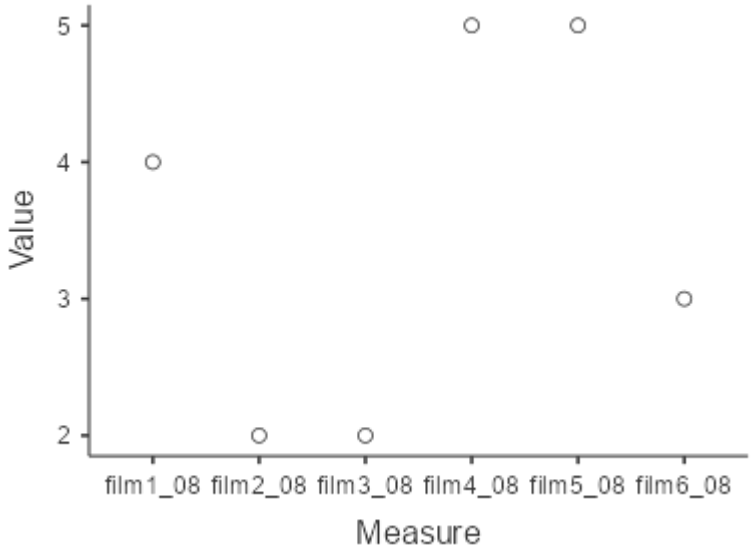
Tabela 44 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania ósmego.

Powyższe obliczenia pokazały, że film 1 (deepfake) różni się od filmu 2, 3 i 6, na co wskazuje $p < 0,001$, natomiast nie zaobserwowano różnicy pomiędzy nim a filmem 4 i 5 ($p > 0,05$). Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4 oraz 5. Nagranie 2 różni się również z nagraniem 6 (deepfake), jednak nie zaobserwowano tej różnicy względem filmów 3, a 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) nie zaobserwowano różnicy. Film 4 i 5 są natomiast różne od filmu 6.

Wnioskować można, iż wpływ filmu 1 (deepfake) na decyzje inwestycje innych osób oglądających go, nie jest istotnie różny od wpływu filmu 4 i 5. Świadczyć to może o ich podobnym odebraniu przez respondentów i scharakteryzowaniu jako wpływające na opinie w zbliżony sposób. Celem weryfikacji tej hipotezy, poniżej umieszczono wykresy przedstawiające średnie oraz wykresy przedstawiające mediany. Wynika to z tego, iż rozkłady wyników są znacząco różne od rozkładu normalnego.



Wykres 46 Średnia odpowiedzi dla pytania ósmego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 47 Mediana odpowiedzi dla pytania ósmego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Brak różnicy pomiędzy filmami 1 (deepfake), a 4 i 5 (średnio znani influencerzy) oraz zbliżone wartości średnich i median sugerować może podobny odbiór nagrań przez respondentów. W ich ocenie filmy te mogą w największym stopniu wpłynąć na decyzje inwestycje innych osób oglądających go. Dzieje się tak zwłaszcza w grupie, która zadeklarowała, iż nie rozpoznała nagrań deepfake.

Średnia drugiego z nagrań deepfake (film 6) zbliżona jest do wyników filmów prezentujących nieznane osoby. Z porównania parami Durbin-Conover, wiadomo jednak, że zaobserwowano istotne statystycznie różnice pomiędzy drugim, a szóstym filmem. Różnice ukazane są również w medianie dla wszystkich trzech nagrań. Film szósty (deepfake) jest wyżej oceniany pod względem potencjalnego wpływu na decyzje inwestycje innych osób oglądających go.

Z przeprowadzonych testów wewnątrzgrupowych Kruskal-Wallis wnioskować można, grupa deklarująca rozpoznanie deepfake oceniała potencjał wpływu nagrań deepfake na decyzje inwestycje innych osób oglądających go, istotnie różnie od osób deklarujących nierozpoznanie deepfake. Wartości median w pierwszym przypadku wynosiły Me1 = 2,00 oraz Me6 = 1,0, natomiast w przypadku grupy deklarującej nierozpoznanie fałszu Me1 = 6,00 a Me6 = 3,5. Pierwsza z nich oceniła możliwość wpływu nagrań deepfake na decyzje inwestycje innych osób oglądających go zdecydowanie niżej niż druga, w której zarówno mediana jak i średnia przewyższała wartości dla filmów prezentujących średnio znanych influencerów.

5.6 Skłonność do działania innych osób

W poniższym podrozdziale przenalizowane zostały aspekty związane z potencjalnym wpływem na działanie (inwestowanie) innych osób niż respondenci. Zastanawiające jest z jakim prawdopodobieństwem, respondenci oceniają możliwość skłonienia innych oglądających je osób do inwestycji.

Pytanie dziewiąte, jakie zadano respondentom w ankiecie, brzmiało: „w jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?”. Odpowiedź udzielano na skali 1 do 10, gdzie 1 to w ogóle, 10 bardzo.

Tym samym było ono w swoim sformułowaniu zbieżne, z pytaniem ósmym („w jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycje innych osób oglądających go?”) jednak miało na celu podkreślenie wpływu nagrań na inne osoby i ich działanie (inwestycję) bezpośrednio po ich obejrzeniu. Pytanie ósme miało więc na celu zbadać, czy nagrania mogą wpływać na decyzje inwestycyjne, natomiast dziewiąte skupia się na potencjale bezpośredniej inwestycji na omawianej na nagraniu platformie.

Założono, iż wyniki w niniejszym podrozdziale będą zbieżne w wynikami otrzymanymi we wcześniejszej analizie. Dla filmów prawdziwych średnio znanych

influencerów wynik ten będzie najwyższy, następnie nieznacznie niżej ocenione będą nagrania deepfake. Najniżej oceniana zachęta do bezpośredniego działania wystąpi w przypadku nagrań prezentujących nieznane osoby.

Statystyki opisowe

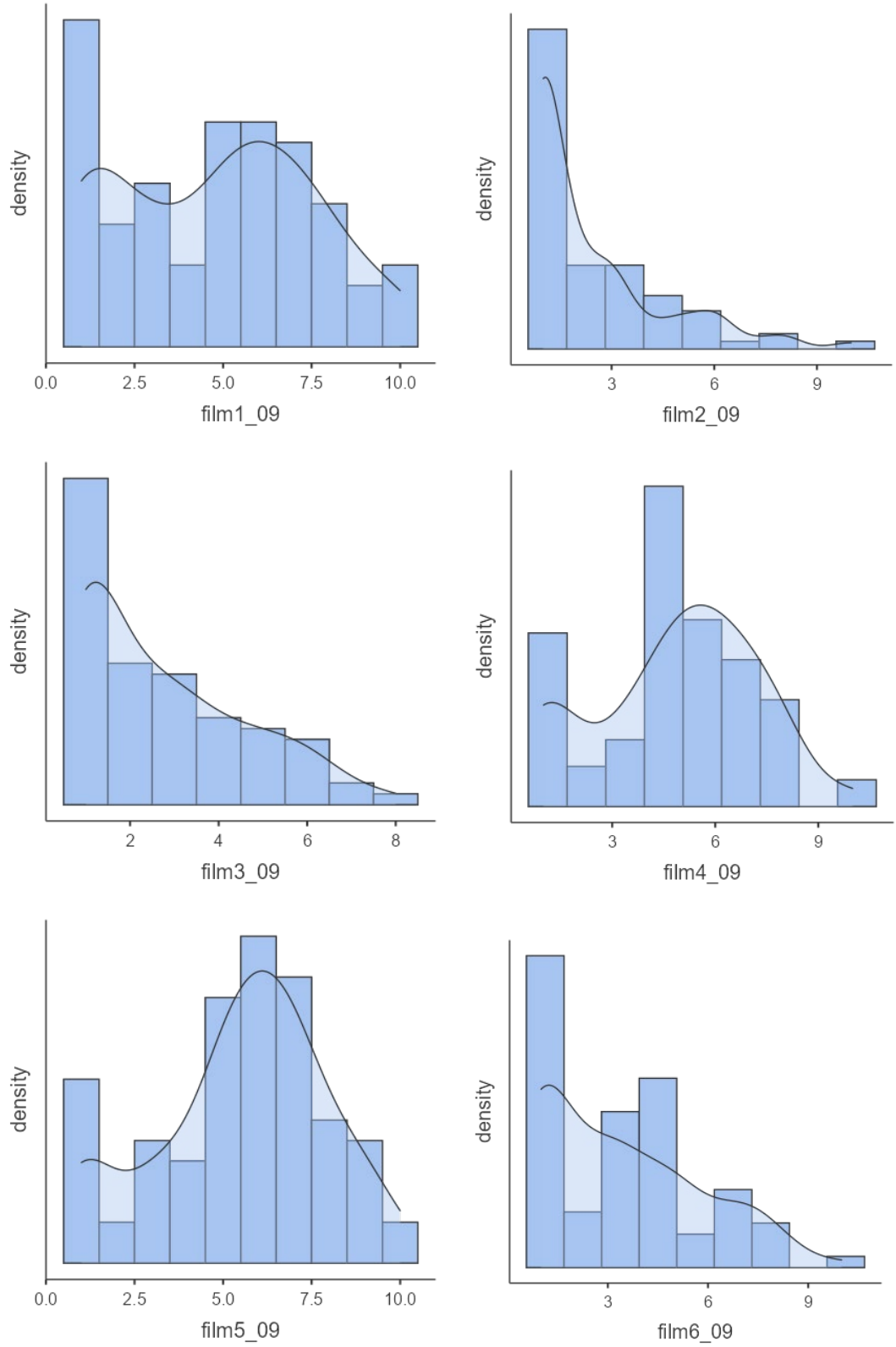
	film1_09	**film2_09**	**film3_09**	**film4_09**	**film5_09**	**film6_09**
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	4.78	2.40	2.75	4.88	5.50	3.38
SE	0.309	0.228	0.211	0.265	0.267	0.267
95% CI dolna granica przedziału ufności dla średniej	4.17	1.95	2.33	4.36	4.98	2.86
95% CI górna granica przedziału ufności dla średniej	5.38	2.85	3.16	5.39	6.02	3.90
Me	5.00	1.00	2.00	5.00	6.00	3.00
D	1.00	1.00	1.00	5.00	6.00	1.00
SD	2.76	2.04	1.88	2.37	2.39	2.38
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	10.0	8.00	10.0	10.0	10.0
SKE	0.0777	1.66	0.867	-0.223	-0.398	0.709
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-1.10	2.35	-0.253	-0.686	-0.482	-0.483
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.928	0.729	0.848	0.935	0.943	0.871
PS-W	< .001	< .001	< .001	< .001	0.001	< .001

Tabela 45 Statystyki opisowe dla dziewiątego pytania „W jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?”.

Dane z powyższej tabeli jednoznacznie pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które za każdym razem przyjmują wartość $p \leqslant 0{,}001$, wartości bezwzględne skośności, które w przypadkach większości filmów są większe od wartości błędu standardowego skośności (poza filmem pierwszym i czwartym) oraz wartości bezwzględne kurtozy, które w połowie przypadków są większe od wartości błędu standardowego tej miary (nagrania 3, 5 i 6). W związku z brakiem rozkładu normalnego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Na poniższych wykresach znajdują się histogramy otrzymanych wyników – dla każdego filmu osobno. Pytanie dziewiąte w kolejności ankiety brzmiało: „w jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków

po obejrzeniu tego nagrania?”. Histogramy te również pokazują, iż rozkład wyników dla żadnego z filmów nie jest zbliżony do rozkładu normalnego.



Wykres 48 Zbiór 6 wykresów odpowiedzi na dziewiąte pytanie dla każdego z sześciu filmów.

Zaprezentowane powyżej wykresy charakteryzuje wysoka różnorodność. Wykresy dla filmu 2 i 3 oraz 6 cechuje silna prawoskośność, zaś dla nagrań 4 i 5 jest to ułożenie jednomodalne, z dominacją średnich wartości. Film 1 (deepfake) odznacza się największą różnorodnością odpowiedzi i przyjmuje wielomodalny rozkład, swoim wyglądem zbliżony do rozkładu płaskiego. Na wykresach dla filmów 2, 3 i 6 widzimy przewagę niskich odpowiedzi, z przewagą ocen filmu 1, podobnie jak w pierwszym filmie.

Celem weryfikacji czy odpowiedzi na pytanie dotyczące skłonności do inwestowania swoich środków po obejrzeniu nagrania, zależy od tego czy respondent rozpoznał czy nie rozpoznał nagrania deepfake, zdecydowano się na dalszą analizę z uwzględnieniem zmiennej nominalnej E, która brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Poniższa tabela przedstawia więc analogiczne dane jak poprzednia, natomiast w tym przypadku rozkłady zostały podzielone względem zmiennej nominalnej E.

Statystyki opisowe

	zmienna nominalna E	**film1_09**	**film2_09**	**film3_09**	**film4_09**	**film5_09**	**film6_09**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	6.00	2.25	2.88	5.28	5.78	4.13
	tak	3.76	2.88	2.60	4.40	4.92	2.52
SE	nie	0.458	0.266	0.341	0.419	0.364	0.416
	tak	0.561	0.536	0.346	0.583	0.547	0.448
95% CI dolna granica przedziału ufności dla średniej	nie	5.10	1.73	2.21	4.46	5.07	3.31
	tak	2.66	1.83	1.92	3.26	3.85	1.64
95% CI górna granica przedziału ufności dla średniej	nie	6.90	2.77	3.54	6.10	6.49	4.94
	tak	4.86	3.93	3.28	5.54	5.99	3.40
Me	nie	7.00	2.00	2.50	6.00	6.00	4.50

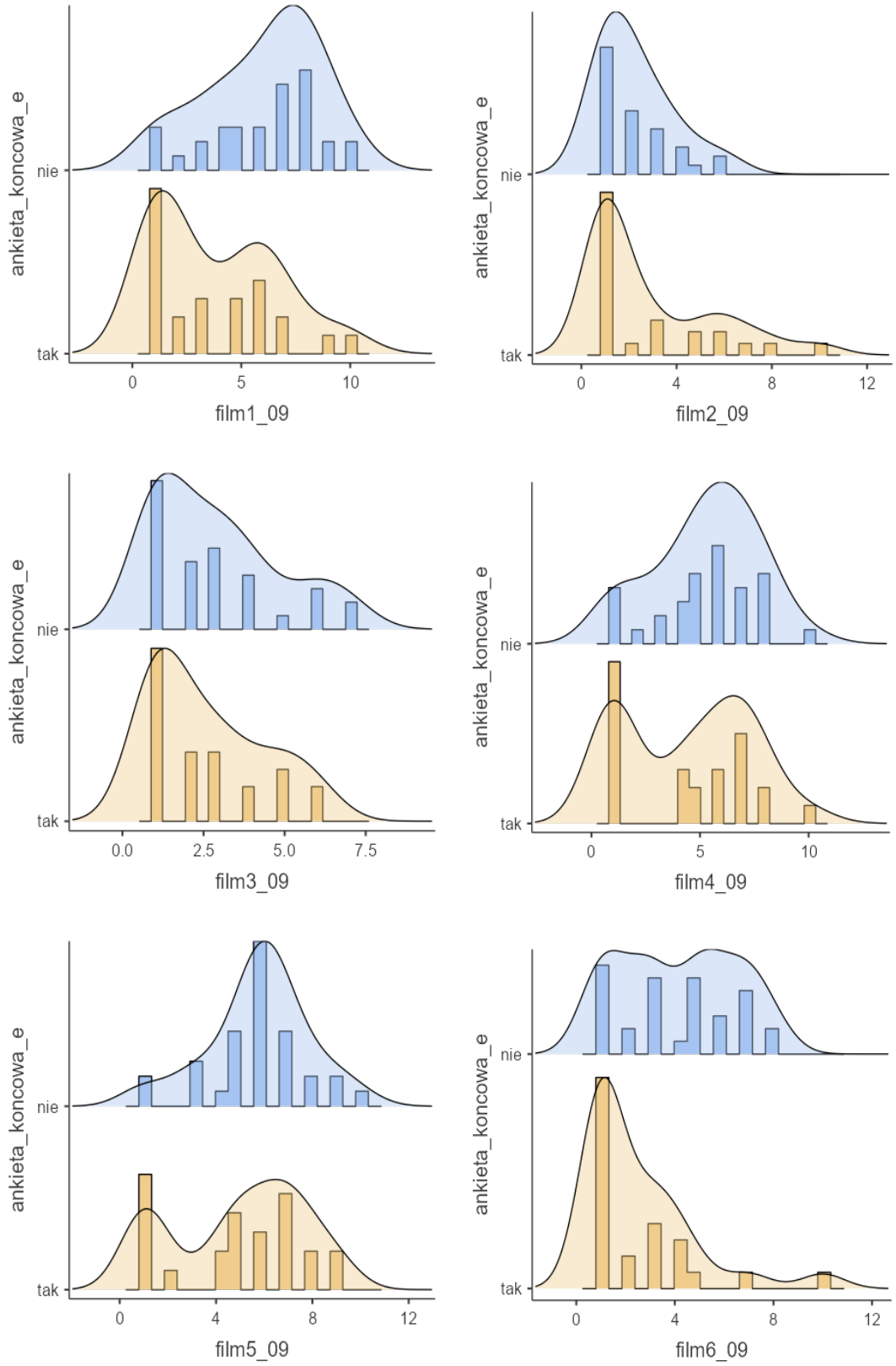
Statystyki opisowe

	zmienna nominalna E	film1_09	film2_09	film3_09	film4_09	film5_09	film6_09
	tak	3.00	1.00	2.00	5.00	5.00	1.00
D	nie	8.00	1.00	1.00	6.00	6.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.59	1.50	1.93	2.37	2.06	2.35
	tak	2.80	2.68	1.73	2.92	2.74	2.24
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	6.00	7.00	10.0	10.0	8.00
	tak	10.0	10.0	6.00	10.0	9.00	10.0
SKE	nie	-0.546	1.19	0.820	-0.389	-0.466	0.0466
	tak	0.625	1.29	0.732	0.0285	-0.320	1.97
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.586	0.653	-0.406	-0.436	0.665	-1.33
	tak	-0.692	0.672	-0.794	-1.36	-1.20	4.38
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.931	0.803	0.855	0.940	0.938	0.908
	tak	0.864	0.744	0.836	0.864	0.890	0.724
PS-W	nie	0.042	< .001	< .001	0.074	0.064	0.010
	tak	0.003	< .001	< .001	0.003	0.011	< .001

Tabela 46 Statystyki opisowe dla dziewiątego pytania „W jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Powyższa tabela przedstawia statystki opisowe dotyczące odpowiedzi na dziewiąte (w kolejności ich prezentacji) pytanie: „w jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?". Dane z powyższej tabeli pokazują, iż nie można przyjąć, iż rozkład uzyskanych wyników jest zbliżony do rozkładu normalnego. Świadczą o tym wartości testu Shapiro-Wilk, które w większości przypadków przyjmują wartość $p < 0,05$ (poza grupą, która nie rozpoznała deepfake dla filmu 4 i 5). Świadczą o tym również wartości bezwzględne skośności, które w przypadku większości filmów są większe od wartości błędu standardowego skośności (poza wartościami skośności dla nagrań 4 i 5). W związku z brakiem rozkładu normlanego uzyskanych wyników, w kolejnych prowadzonych analizach statystycznych, zastosowano testy nieparametryczne.

Poniżej znajdują się histogramy dla każdego filmu, przedstawiające wyniki badanych osób, które zostały podzielone według zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”). Histogramy te ilustrują, jak rozkładały się odpowiedzi w zależności od tego, czy uczestnicy badania uważali, że rozpoznali fałszywe nagrania.



Wykres 49 Zbiór 6 wykresów odpowiedzi na dziewiąte pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na pytanie dziewiąte, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmów 1 i 6 (deepfake) zauważyć można zdecydowaną przewagę niskich odpowiedzi, dla grupy która rozpoznała, iż jest to nagranie fałszywe (dominacja odpowiedzi 1). Zaskakująca jest jednak również rozbieżność w ułożeniu wyników dla nagrań 4 i 5, gdzie w grupie deklarującej rozpoznanie deepfake również dominuje odpowiedź 1, a wykres z ułożenia jednomodalnego, przybiera kształt bimodalny.

Aby sprawdzić, czy różnice między grupami są istotne statystycznie, przeprowadzono serię testów Kruskal-Wallis. Wyniki tych testów zostały przedstawione w poniższej tabeli. Z analizy wynika, że jedynie w przypadku pierwszego oraz ostatniego filmu (deepfake) zaobserwowano istotne różnice w rozkładach wyników ($p < 0,05$). Oznacza to, że w tych dwóch przypadkach różnice między grupami są na tyle znaczące, że mogą być uznane za statystycznie istotne. Odrzuca to hipotezę o istności statystycznej różnic w układzie odpowiedzi dla nagrań 4 i 5. Świadczy to o tym, iż różnice pomiędzy tymi dwoma grupami nie są wystarczająco duże, aby móc je wiarygodnie wyjaśnić przy użyciu statystyk. Istotność statystyczna została użyta w celu potwierdzenia czy różnice między grupami są na tyle duże, aby nie można było ich uznać za przypadkowe, co oznacza, że mogą być uznane za prawdziwe w sensie statystycznym. Zaobserwowano to jedynie w przypadku filmów deepfake – pierwszy i szósty. W tych dwóch przypadkach różnice między grupami są statystycznie istotne, co sugeruje, że oceny respondentów znacznie się różniły w zależności od tego, czy rozpoznali fałszywość nagrań.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_09	8.81900	1	0.003	0.15748
film2_09	0.00266	1	0.959	4.75e-5
film3_09	0.25629	1	0.613	0.00458
film4_09	1.20962	1	0.271	0.02160
film5_09	0.92582	1	0.336	0.01653
film6_09	7.47390	1	0.006	0.13346

Tabela 47 Test Kruskal-Wallis dla odpowiedzi do pytania dziewiątego.

W celu przetestowania hipotezy mówiącej o tym, że odpowiedzi na dziewiąte pytanie („w jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?") różnią się w zależności od filmu,

przeprowadzono test Friedmana. Wyniki testu jednoznacznie wskazują, że oceny wszystkich filmów, zarówno prawdziwych, jak i fałszywych, różnią się istotnie statystycznie między sobą (χ^2 (5) = 128; p < 0,001). Aby dokładniej zbadać różnice między odpowiedziami dotyczącymi poszczególnych filmów, przeprowadzono test porównania parami – Durbin-Conover, którego wyniki zostały przedstawione w poniższej tabeli.

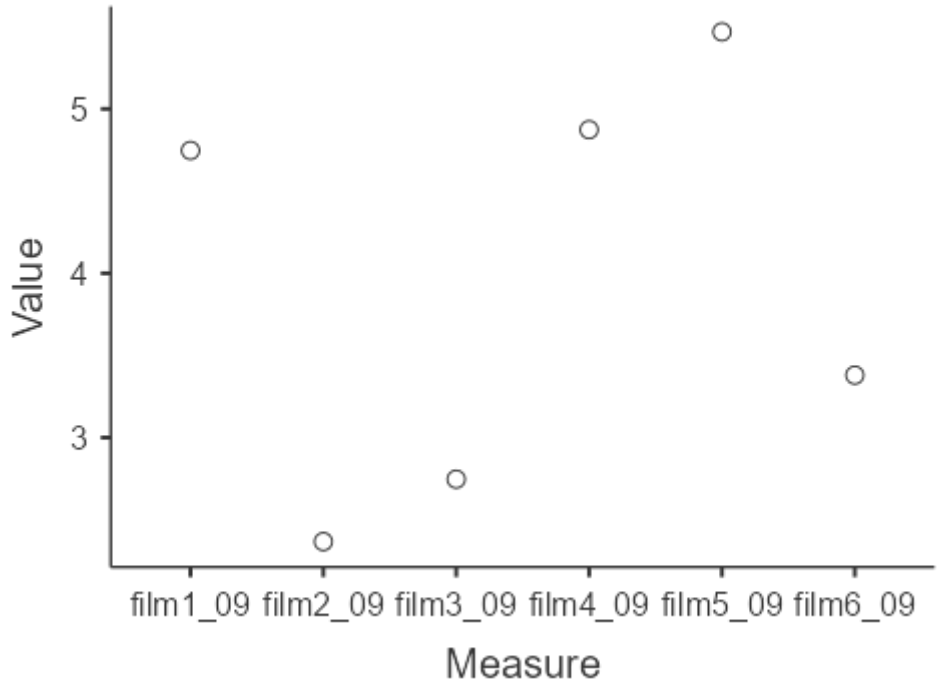
Porównania Parami (Durbin-Conover)

			Statistic	p
film1_09	-	film2_09	8.900	< .001
film1_09	-	film3_09	7.388	< .001
film1_09	-	film4_09	0.371	0.711
film1_09	-	film5_09	1.569	0.117
film1_09	-	film6_09	4.764	< .001
film2_09	-	film3_09	1.512	0.131
film2_09	-	film4_09	8.530	< .001
film2_09	-	film5_09	10.469	< .001
film2_09	-	film6_09	4.136	< .001
film3_09	-	film4_09	7.018	< .001
film3_09	-	film5_09	8.957	< .001
film3_09	-	film6_09	2.624	0.009
film4_09	-	film5_09	1.940	0.053
film4_09	-	film6_09	4.393	< .001
film5_09	-	film6_09	6.333	< .001

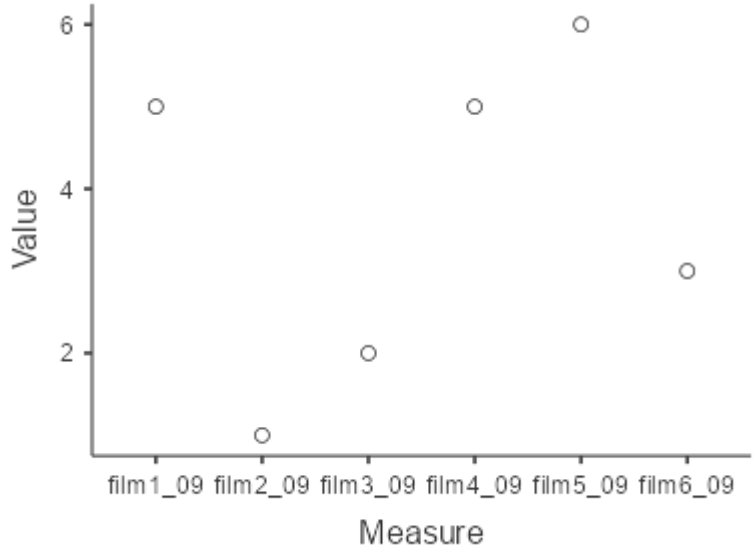
Tabela 48 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania dziewiątego.

Powyższe obliczenia pokazały, iż film 1 (deepfake) różni się od filmów 2, 3 i 6 (p<0,05), natomiast nie zaobserwowano różnic pomiędzy nim a filmem 4 i 5 (p>0,05). Pomiędzy nagraniami 2, a 3 nie zaobserwowano różnic. Różnica w podejrzeniu skłonności do zainwestowania pieniędzy po obejrzeniu nagrań jest jednak różna pomiędzy nimi, a filmami 4, 5 i 6. Nie stwierdzono również różnicy pomiędzy nagraniami 4, a 5. Oba są natomiast różne od filmu 6 (deepfake).

Z porównania Durbin-Conover wnioskować można, iż zdaniem respondentów, pierwszy film deepfake mógł w podobnym stopniu wpłynąć na podejrzenie skłonności do zainwestowania pieniędzy po jego obejrzeniu, co film czwarty i piąty (średnio znani influencerzy). Celem weryfikacji tych hipotez, poniżej umieszczono wykres przedstawiający średnie oraz wykres przedstawiający mediany.



Wykres 50 Średnia odpowiedzi dla pytania dziewiątego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 51 Mediana odpowiedzi dla pytania dziewiątego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Zarówno średnia, jak i media ukazują podobieństwo w odbiorze pierwszego filmu deepfake, jak i nagrań średnio znanych influencerów. Respondenci nadali tym trzem filmom charakter najsilniej przekonywujących do inwestycji. Nieznacznie gorzej oceniony został drugi z filmów deepfake, jednak z porównania Durbin-Conover wiadomo, iż jest on statystycznie różny we wpływie od pozostałych nagrań. Zgodnie

z założeniem przedstawionym w niniejszym podrozdziale, najniższy wynik otrzymały filmy z nieznanymi osobami. Na podstawie porównania parami wnioskować można, iż nawet słabszej jakości deepfake jest zdaniem respondentów w większym stopniu skłonić do inwestycji niż nagrania prezentujące nieznane osoby.

5.7 Wnioski

U podstaw bezpieczeństwa narodowego leży bezpieczeństwo personalne. Pojęcie odnosi się do ochrony jednostek przed krzywdą, taką jak przestępstwa, przemoc, wypadki lub choroby. W bezpieczeństwie narodowym jednostką będzie obywatel, zaś zagrożeniem omawianym w niniejszym rozdziale jest wpływ nagrań deepfake. W kontekście przeprowadzonego eksperymentu (oszustwo inwestycyjne) bezpośrednio sprawdzano wpływ nagrań na decyzje inwestycyjne jednostki. Pośrednio można jednak wyciągnąć wnioski dotyczące szerszego problemu manipulowania internautą przy pomocy fałszywych nagrań.

Pytanie badawcze, postawione w niniejszym rozdziale, brzmiało: jaki wpływ ma percepcja zmanipulowanych nagrań wideo na wybory i opinie internautów, w świetle bezpieczeństwa narodowego? Hipoteza, postawiona przed dokonanym badaniem, zakładała zaś, iż percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów, a jej oddziaływanie na bezpieczeństwo narodowe jest zależne od stopnia wiarygodności tych nagrań oraz od kontekstu, w jakim są one prezentowane.

W toku analizy zgromadzonego materiału empirycznego udało się zweryfikować tę hipotezę – percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów. Analizie poddano odpowiedzi na pytania potencjale nagrań deepfake do przekonywania i wpływu na osoby je oglądające, możliwości wykorzystania technologii do oszustw inwestycyjnych, wpływ nagrań na decyzje podejmowania ryzyka oraz subiektywne odczucia potencjału wpływu nagrań na inne osoby. Kompleksowa analiza pozwoliła zweryfikować stawianą hipotezę i odpowiedzieć na zadane pytanie badawcze.

Skuteczność w przekonywaniu do fałszywych inwestycji nagrań deepfake, nie jest na tyle duża co filmów prezentujących prawdziwych influencerów. Jest ona jednak niewiele mniejsza, a znacznie większa od filmów prezentujących nieznane osoby. Ponadto, co istotne, same nagrania zachęcające do inwestycji (zarówno prawdziwe jak

i deepfake) w niewielkim stopniu wpływają na ilość kapitału, który oszukiwana osoba jest skora zainwestować. Podobnie dzieje się w przypadku przekonania co do realności zysku. Osoby badane wskazały, iż odczucie realności obiecywanego zysku po obejrzeniu pierwszego filmu deepfake jest dla zbliżone do tego po obejrzeniu prawdziwych nagrań średnio znanych influencerów. Oba nagrania deepfake oceniane są znacznie wyżej i są statystycznie różne od filmów prezentujących nieznane osoby.

Co istotne, nie ma istotnych statystycznie różnic pomiędzy pierwszym nagraniem, a filmami średnio znanych influencerów pod względem potencjalnego wpływu filmów deepfake na inne osoby. Dzieje się tak zarówno w przypadku pytania o decyzje inwestycje innych osób oglądających go jak i odczucia ich skłonności do zainwestowania pieniędzy po obejrzeniu nagrania. Tym samym osoby badane uważają, iż one same nie dałyby się przekonać do fałszywej inwestycji w takim samym stopniu jak inne osoby. Przejawiać się tu może efekt nadmiernej pewności siebie lub efekt Dunninga-Krugera. Wątek ten rozwinięty został w podrozdziale 6.4.

Deepfake to ciągle rozwijająca się technologia, która pozwala na tworzenie wiarygodnie wyglądających fałszywych treści wideo. Coraz częściej nagrania deepfake są używane w celu manipulowania informacjami na temat kandydatów politycznych, przedsiębiorstw lub produktów finansowych. Może to prowadzić do wprowadzenia w błąd opinii publicznej i wpływu na wybory oraz decyzje ekonomiczne. Ponadto, biorąc pod uwagę bezpieczeństwo ekonomiczne, wnioskiem z badania jest również to, iż wpływ percepcji zmanipulowanych nagrań wideo na bezpieczeństwo narodowe zależy od stopnia wiarygodności tych nagrań oraz kontekstu, w jakim są one prezentowane. Jeżeli nagrania deepfake są używane w celu rozpowszechniania fałszywych informacji na temat ważnych kwestii gospodarczych, takich jak decyzje polityczne, prognozy ekonomiczne czy wiadomości o firmach, może to wpłynąć na zaufanie inwestorów, rynek finansowy oraz ogólną stabilność gospodarczą.

Należy przede wszystkim zaznaczyć znaczny wpływ nagrań deepfake na bezpieczeństwo personalne, które jest szczególnie istotne w kontekście ekonomicznym. Filmy deepfake mogą być wykorzystywane w celu oszustw finansowych, takich jak fałszywe inwestycje, fałszywe nagrania rozmów telefonicznych czy wideokonferencji w celu wyłudzenia pieniędzy lub poufnych informacji handlowych. Osoby prywatne, w tym przedsiębiorcy i inwestorzy, mogą stać się ofiarami takich

manipulacji, co może prowadzić do poważnych strat finansowych i naruszenia ich bezpieczeństwa personalnego.

Zasygnalizowane zostały również potrzeby rozwinięcia skutecznych narzędzi i technologii ochrony przed deepfake. To, że deepfake może wpływać na wybory, opinie i bezpieczeństwo narodowe, podkreśla konieczność ciągłego doskonalenia systemów wykrywania deepfake oraz edukacji społeczeństwa na temat zagrożeń związanych z tą technologią. Rządy, instytucje finansowe i firmy powinny również inwestować w odpowiednie procedury, które pomogą weryfikować autentyczność treści wideo oraz chronić swoje interesy i dane.

Rozdział 6. Przeciwdziałanie dezinformacji wyzwaniem dla bezpieczeństwa strukturalnego

Poniższy rozdział koncentruje się na wyszczególnieniu i uszczegółowieniu aspektów wpływających na proliferację dezinformacji w Internecie, szczególnie pod postacią fałszywych nagrań deepfake. W rozdziale przeanalizowano elementy mogące wpływać na percepcję dezinformacji przez jednostkę oraz wpływ nagrań deepfake na jej działania i podejmowane decyzje.

Celem analiz zamieszczonych w niniejszym rozdziale, jest ustalenie czy istnieją cechy, które w znacznym stopniu mogą przyczynić się do wzrostu lub spadku podatności na dezinformację nagraniami deepfake. Pytanie szczegółowe prezentowane w niniejszym rozdziale brzmi: Jaki wpływ mają zastosowane moderatory na odbiór dezinformacji deepfake, a przez to na bezpieczeństwo narodowe? Zaproponowana hipoteza brzmi następująco: osoby z podwyższonymi poszczególnymi wskaźnikami społecznymi są bardziej podatne na uleganie manipulacjom dezinformacji deepfake, co przyczynia się do zwiększenia ryzyka zagrożenia bezpieczeństwa narodowego.

W rozdziale poruszono również wątki dotyczące możliwości przeciwdziałania dezinformacji multimedialnej. Przeanalizowano dotychczas podejmowane działania – ich zasięg, jak i skuteczność. Zaproponowano również autorskie rozwiązania, obejmujące zarówno aspekty indywidualne, jak i w szczególności systemowe. Drugie pytanie badawcze, zaprezentowane w niniejszym rozdziale brzmi: jakie należy podjąć działania w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake? Weryfikowana hipoteza zakłada, iż aby ochronić się przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, należy podjąć działania zapobiegające rozprzestrzenianiu się takich materiałów, edukować społeczeństwo w zakresie rozpoznawania dezinformacji audiowizualnej oraz zapewnić odpowiednie narzędzia do weryfikacji prawdziwości takich materiałów.

W rozdziale holistycznie ujęto problem fałszywych nagrań deepfake w Internecie i możliwości prowadzenia nielegalnych działań z nimi związanymi. Na podstawie analiz częściowych wyników, opracowanych w poprzednich rozdziałach oraz weryfikacji stawianych hipotez, wyciągnięto wnioski, które po zestawieniu ze sobą utworzyły zwartą konkluzję udzielającą odpowiedzi na drugie pytanie badawcze.

6.1 Ocena wpływu nagrań na odbiór aktorów przez ich otoczenie

Nagrania deepfake, nawet te nieprofesjonalne, mogące być łatwo rozpoznane, hipotetycznie mogą wpływać na opinię o osobach, które występują na danym nagraniu. W przypadku nierozpoznania oszustwa, osoba poszkodowana może obwiniać winą za poniesione straty osobę, której wizerunek zachęcał ją do oszukańczej inwestycji. W trakcie lektury zgłoszeń popularnych oszustw internetowych, często odnaleźć można informacje, że oszukana osoba została zachęcona do rejestracji na platformie inwestycyjnej przez Prezes Rady Ministrów, Prezydenta RP, celebrytę lub danego ministra[266]. Osoby poszkodowane mogę wówczas czuć się pokrzywdzone przez osobę, której wizerunek był prezentowany w danej reklamie. Może się to przełożyć na spadek zaufania względem tej osoby oraz inne konsekwencje wizerunkowe.

W odmiennej sytuacji znajdują się osoby, które rozpoznały deepfake. Hipotetycznie mogą one odnieść wrażenie, iż wizerunek prezentowanej osoby oraz kontekst jego wykorzystania ośmiesza ją. Ponadto w obu przypadkach, jeżeli znamy danego polityka czy celebrytę może to wpłynąć na nasz odbiór treści przez niego reklamowanych. Celem sprawdzenia, czy bezprawne wykorzystanie wizerunku influencera do oszustwa, może wpłynąć na postrzeganie go przez jego otoczenie, postanowiono zadać następujące pytanie: „jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?".

Na wstępie przypomnieć należy, iż pierwsze analizowane nagranie było filmem deepfake z twarzą znanej aktorki, która zachęcała do inwestycji na fałszywej platformie. Drugie i trzecie nagranie przedstawiało nieznane osoby również promujące fałszywą inwestycję. Z kolei czwarte i piąte wideo zawierało średnio popularnych influencerów reklamujących tę samą ofertę. Wideo szóste, podobnie jak pierwsze, było filmem wytworzonym z wykorzystaniem technologii deepfake, przedstawiającym średnio znanego influencera promującego oszustwo. Odpowiedzi na pytania dotyczące filmów były oceniane w skali od 1 (w ogóle) do 10 (bardzo). Poniżej znajdują się statystyki opisowe dotyczące dwunastego pytania.

[266] Komunikat Policji opisujący zgłoszenie utraty oszczędności przez 77-letnią kobietę. https://opolska.policja.gov.pl/op/aktualnosci/115700,77-latka-chciala-zainwestowac-pieniadze-w-Internecie-stracila-270-tysiecy-zlotyc.html [dostęp: 01.01.2024].

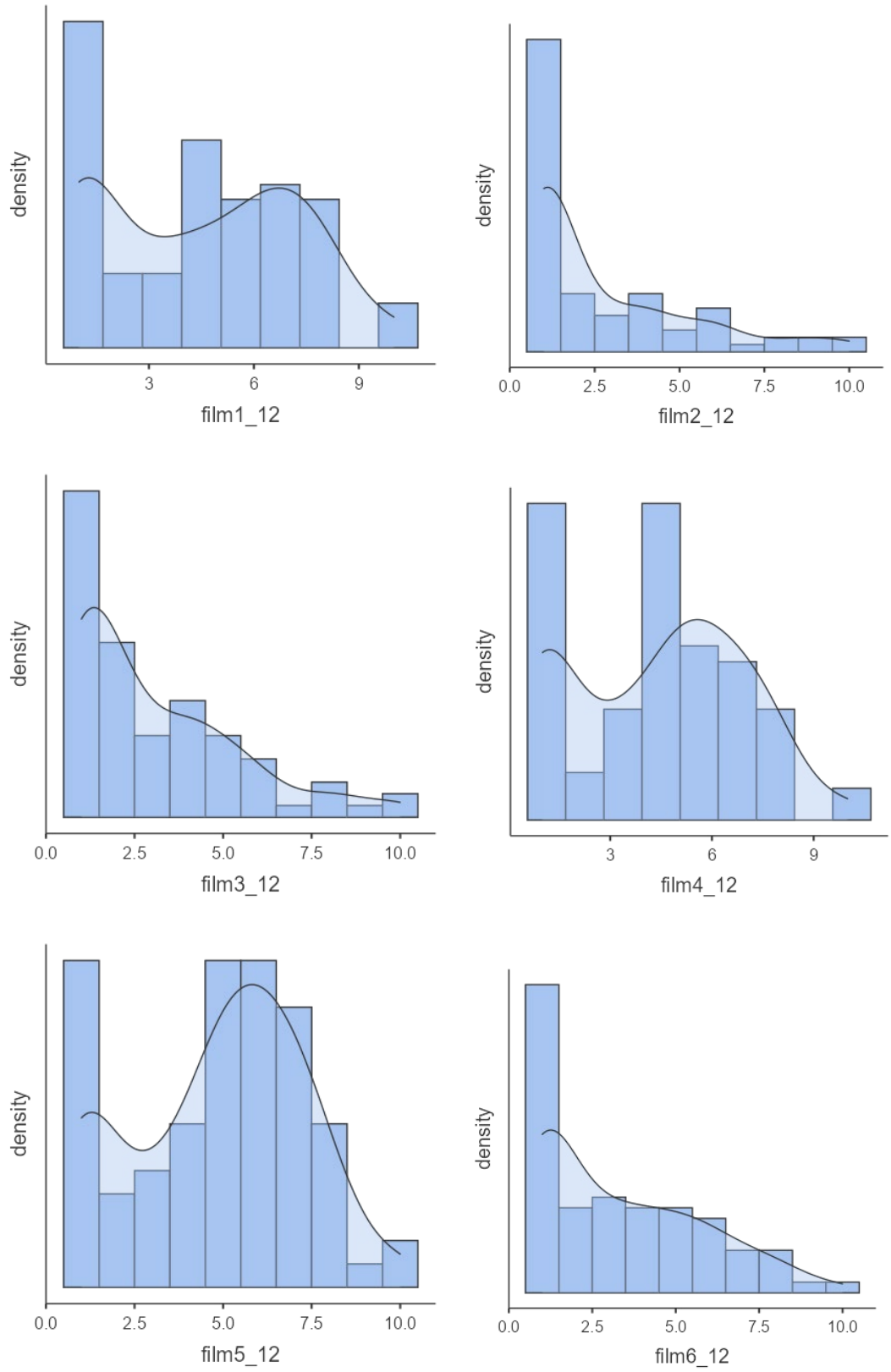
Statystyki opisowe

	film1_12	film2_12	film3_12	film4_12	film5_12	film6_12
N	80	80	79	80	80	79
Brakujące odpowiedzi	2	2	3	2	2	3
M	4.45	2.73	3.09	4.41	4.85	3.35
SE	0.313	0.277	0.266	0.286	0.274	0.275
95% CI dolna granica przedziału ufności dla średniej	3.84	2.18	2.57	3.85	4.31	2.82
95% CI górna granica przedziału ufności dla średniej	5.06	3.27	3.61	4.97	5.39	3.89
Me	4.50	1.00	2.00	5.00	5.00	3.00
D	1.00	1.00	1.00	1.00	1.00 [a]	1.00
SD	2.80	2.48	2.36	2.56	2.46	2.44
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	10.0	10.0	10.0	10.0	10.0
SKE	0.111	1.43	1.19	0.0272	-0.196	0.754
SEk	0.269	0.269	0.271	0.269	0.269	0.271
K	-1.28	1.16	0.800	-1.06	-0.834	-0.460
Std. error K	0.532	0.532	0.535	0.532	0.532	0.535
S-W	0.895	0.735	0.831	0.914	0.931	0.863
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

[a] Istnieje więcej niż jedno D, tylko pierwsze jest odnotowywane.

Tabela 49 Statystyki opisowe dla dwunastego pytania „Jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?”.

Dane z powyższej tabeli jednoznacznie wskazują, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Potwierdzają to wartości testu Shapiro-Wilk, które we wszystkich przypadkach są mniejsze niż 0,001 ($p < 0,001$). Dodatkowo, wartości skośności dla większości filmów przekraczają wartość błędu standardowego tej miary, a w dwóch przypadkach są one większe niż 1, co wskazuje na asymetryczność rozkładu. Wartości bezwzględne kurtozy są w większości przypadków większe od błędu standardowego, co świadczyć może o większej koncentracji wyników. Z uwagi na brak normalności rozkładu, w dalszych analizach zastosowano testy nieparametryczne, które są bardziej odpowiednie w przypadku tego typu danych.



Wykres 52 Zbiór 6 wykresów odpowiedzi na dwunaste pytanie dla każdego z sześciu filmów.

Na wykresach 2, 3 i 6 zauważalna jest przewaga niskich odpowiedzi, co wskazuje na silną prawoskośność rozkładów, z wyraźną dominacją odpowiedzi „1 – wcale". Szczególnie jest to widoczne w przypadku filmów drugiego i trzeciego, które prezentowały wizerunki nieznanych osób. Nagrania pierwsze (deepfake), czwarte i piąte zostały ocenione bardziej równomiernie, co może sugerować, że część respondentów dostrzegała wpływ tych nagrań na odbiór wizerunku osób w nich występujących.

Aby dokładniej przeanalizować, jak rozpoznanie deepfake wpłynęło na odpowiedzi, poniżej przedstawiono tabelę zawierającą analogiczne dane jak poprzednia. Tym razem rozkłady wyników zostały podzielone w zależności od odpowiedzi na pytanie nominalne E: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Tabela ta pozwala na zbadanie różnic w ocenach pomiędzy respondentami, którzy uważali, że rozpoznali deepfake, a tymi, którzy tego nie zrobili.

Statystyki opisowe

	zmienna nominalna E	**film1_12**	**film2_12**	**film3_12**	**film4_12**	**film5_12**	**film6_12**
N	nie	32	32	32	32	32	32
	tak	25	25	25	25	25	25
Brakujące odpowiedzi	nie	0	0	0	0	0	0
	tak	0	0	0	0	0	0
M	nie	5.59	2.59	2.97	4.63	5.25	4.03
	tak	3.76	2.92	2.68	4.20	4.56	1.88
SE	nie	0.455	0.464	0.388	0.418	0.389	0.429
	tak	0.533	0.510	0.461	0.608	0.513	0.273
95% CI dolna granica przedziału ufności dla średniej	nie	4.70	1.68	2.21	3.81	4.49	3.19
	tak	2.72	1.92	1.78	3.01	3.55	1.35
95% CI górna granica przedziału ufności dla średniej	nie	6.49	3.50	3.73	5.44	6.01	4.87
	tak	4.80	3.92	3.58	5.39	5.57	2.41
Me	nie	6.00	1.00	2.00	5.00	6.00	4.00
	tak	3.00	1.00	2.00	5.00	5.00	1.00
D	nie	7.00 [a]	1.00	1.00	6.00	6.00	1.00

Statystyki opisowe

	zmienna nominalna E	film1_12	film2_12	film3_12	film4_12	film5_12	film6_12
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.58	2.63	2.19	2.37	2.20	2.43
	tak	2.67	2.55	2.30	3.04	2.57	1.36
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	10.0	10.0	10.0	10.0	9.00
	tak	10.0	10.0	10.0	10.0	10.0	5.00
SKE	nie	-0.713	1.66	1.51	-0.128	-0.378	0.148
	tak	0.553	1.30	1.63	0.189	-0.0431	1.31
SEk	nie	0.414	0.414	0.414	0.414	0.414	0.414
	tak	0.464	0.464	0.464	0.464	0.464	0.464
K	nie	-0.566	1.70	2.51	-0.532	-0.109	-1.16
	tak	-0.733	1.13	2.81	-1.48	-0.726	0.364
Std. error K	nie	0.809	0.809	0.809	0.809	0.809	0.809
	tak	0.902	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.876	0.671	0.826	0.926	0.936	0.913
	tak	0.876	0.779	0.760	0.838	0.917	0.687
PS-W	nie	0.002	< .001	< .001	0.031	0.058	0.013
	tak	0.006	< .001	< .001	0.001	0.043	< .001

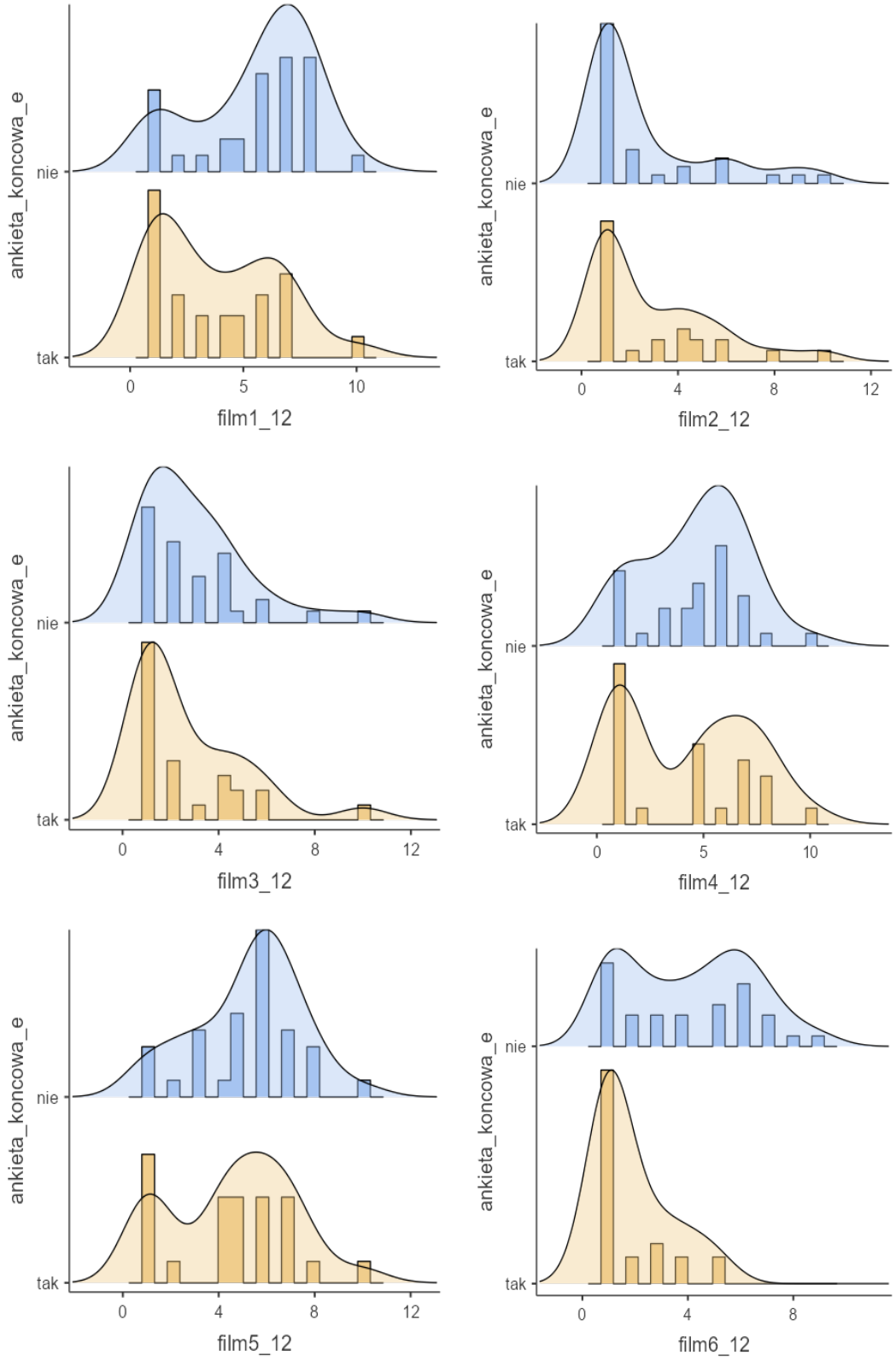
[a] Istnieje więcej niż jedno D, tylko pierwsze jest odnotowywane.

Tabela 50 Statystyki opisowe dla dwunastego pytania „Jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Podobnie jak w poprzedniej tabeli, wszystkie rozkłady wyników są różne od normalnego. Wskazują na to wartości kurtozy oraz skośności, gdzie w połowie analizowanych przypadków wartość bezwzględna tych miar jest wyższa niż wartość odpowiadających im błędów standardowych. Świadczą o tym również niskie wyniki (w większości $p < 0,05$, z wyjątkiem nagrania piątego, grupa odpowiadająca „nie") testu Shapiro-Wilk.

Poniżej przedstawiono histogramy dla każdego filmu, które ilustrują rozkład odpowiedzi w zależności od zmiennej nominalnej E, brzmiącej: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Histogramy te umożliwiają porównanie ocen pomiędzy grupami respondentów, którzy deklarowali rozpoznanie deepfake oraz tych, którzy tego nie zrobili. Analiza tych wykresów pozwala na dalsze zrozumienie, jak świadomość

obecności deepfake wpływa na postrzeganie treści wideo oraz ich potencjalny wpływ na odbiór danej osoby.



Wykres 53 Zbiór 6 wykresów odpowiedzi na dwunaste pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela wraz z powyższymi wykresami histogramów obrazują różnice w odpowiedzi na dwunaste pytanie, między osobami, które uważają, iż udało im się rozpoznać deepfake, a tymi, które uważają, że tego nie zrobiły. Zwłaszcza dla filmu 6 (deepfake) zauważyć można zdecydowaną przewagę wysokich odpowiedzi, dla grupy która nie rozpoznała, iż jest to nagranie fałszywe. Zdaniem dużej liczby respondentów, którzy twierdzą, że rozpoznali deepfake, wideo to nie wpływa na wizerunek osoby w nim występującej (znaczna dominacja odpowiedzi 1 – „w ogóle"). Podobną zależność zaobserwować można w przypadku pierwszego filmu (również deepfake). Jednak ze względu na lepsze wykonanie wideo oraz mniejszą liczbę rozpoznania fałszu, odpowiedzi te nie kontrastują ze sobą tak intensywnie jak w przypadku filmu szóstego.

Aby ocenić, czy różnice w odpowiedziach są istotne statystycznie, przeprowadzono serię nieparametrycznych testów Kruskal-Wallis, będących odpowiednikiem jednoczynnikowej analizy wariancji (ANOVA). Wyniki tych testów przedstawiono w poniższej tabeli. Analiza wykazała, że istotne różnice statystyczne występują jedynie w przypadku pierwszego oraz ostatniego filmu, które oba są nagraniami deepfake ($p < 0,05$). Oznacza to, że różnice pomiędzy grupami respondentów, które rozpoznały deepfake i które go nie rozpoznały, są na tyle znaczące, aby można je było interpretować jako rzeczywiste, a nie przypadkowe. Potwierdzenie istotności statystycznej umożliwia dalsze analizy i wnioski dotyczące wpływu rozpoznania deepfake na oceny wideo.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_12	6.711	1	0.010	0.11984
film2_12	0.375	1	0.540	0.00670
film3_12	0.709	1	0.400	0.01266
film4_12	0.225	1	0.635	0.00402
film5_12	1.175	1	0.278	0.02098
film6_12	12.447	1	< .001	0.22228

Tabela 51 Test Kruskal-Wallis dla odpowiedzi do pytania dwunastego.

W celu zweryfikowania hipotezy dotyczącej wpływu różnych filmów na odbiór osoby przez znajomych lub obserwujących, przeprowadzono test Friedmana. Wyniki tego testu wskazały, że odpowiedzi na dwunaste pytanie różnią się istotnie pomiędzy wszystkimi filmami, zarówno prawdziwymi, jak i deepfake (χ^2 (5) = 65,3; $p < 0,001$).

Aby dokładniej zbadać różnice pomiędzy poszczególnymi odpowiedziami na to pytanie, zastosowano nieparametryczny odpowiednik testów post hoc. W tym przypadku porównania przeprowadzono parami, wykorzystując test Durbin-Conover. Tabela z wynikami tego testu znajduje się poniżej, co umożliwia szczegółową analizę różnic w ocenach wpływu nagrań na postrzeganie osób w nich występujących.

Porównania Parami (Durbin-Conover)

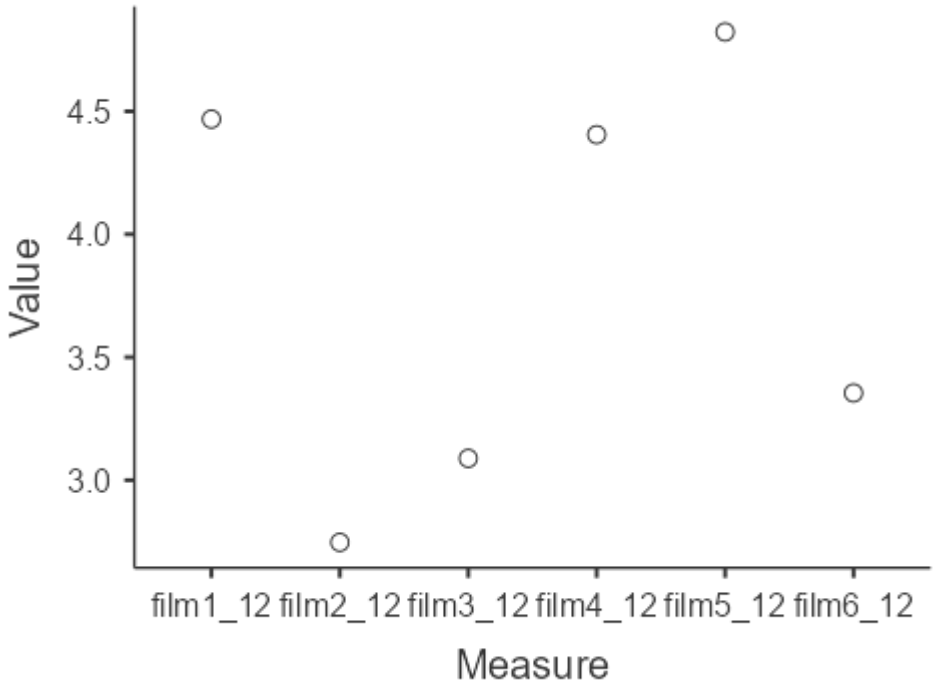
			Statistic	p
film1_12	-	film2_12	6.028	< .001
film1_12	-	film3_12	4.744	< .001
film1_12	-	film4_12	0.839	0.402
film1_12	-	film5_12	0.655	0.513
film1_12	-	film6_12	3.984	< .001
film2_12	-	film3_12	1.284	0.200
film2_12	-	film4_12	5.189	< .001
film2_12	-	film5_12	6.683	< .001
film2_12	-	film6_12	2.044	0.042
film3_12	-	film4_12	3.905	< .001
film3_12	-	film5_12	5.399	< .001
film3_12	-	film6_12	0.760	0.448
film4_12	-	film5_12	1.494	0.136
film4_12	-	film6_12	3.145	0.002
film5_12	-	film6_12	4.639	< .001

Tabela 52 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania dwunastego.

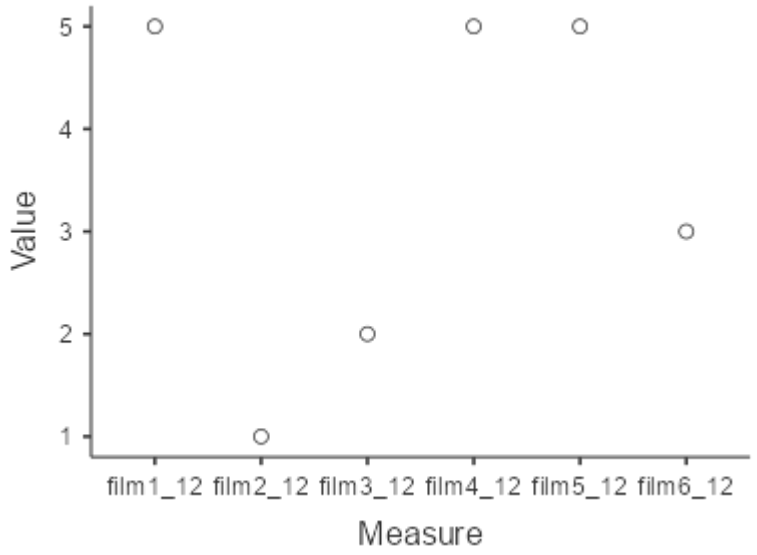
Powyższe obliczenia pokazały, że względem pytania dwunastego, film 1 różni się od filmu 2, 3 i 6, na co wskazuje $p < 0,001$, natomiast nie zaobserwowano różnicy pomiędzy nim a filmem 4 i 5 ($p > 0,05$). Pomiędzy nagraniami 2, a 3 nie zanotowano różnic (oba prezentowały nieznane osoby), natomiast różnica jest między nimi, a nagraniami 4 oraz 5. Zastanawiająca jest różnica pomiędzy nagraniem 3, a 6 ($p < 0,05$), której nie zaobserwowano w zestawieniu nagrań 2, a 6. Z porównania Durbin-Conover wiemy, że pomiędzy nagraniami 4, a 5 (prawdziwe nagrania influencerów) również nie odnotowano istotnej statystycznie różnicy. Film 4 i 5 są natomiast różne od filmu 6.

Z analizy wyników wynika, że film pierwszy nie wykazuje istotnych różnic w odbiorze osoby przez znajomych, osoby obserwujące ją w porównaniu z filmami 4 i 5, co sugeruje, że widzowie odbierają ocenę w sposób zbliżony. Aby dokładniej zweryfikować tę hipotezę, poniżej przedstawiono wykresy ilustrujące zarówno średnie, jak i mediany odpowiedzi. Wybór prezentacji obu miar jest uzasadniony, ponieważ rozkłady wyników znacząco odbiegają od normalności. W takich przypadkach średnie

mogą być mocno zniekształcone przez wartości odstające, co sprawia, że mediana stanowi bardziej adekwatną miarę tendencji centralnej i lepiej oddaje rzeczywisty rozkład ocen widzów.



Wykres 54 Średnia odpowiedzi dla pytania dwunastego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 55 Mediana odpowiedzi dla pytania dwunastego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Mediana pokazuje, iż filmy 2 oraz 3 i 6 (deepfake) zdaniem respondentów, nie mają aż tak znacznego wpływu na ocenę osób w nich występujących, jak w przypadku

filmów 1, 4 i 5. Respondenci zdają się tym samym uważać, iż w przypadku bardziej rozpoznawalnych osób, potencjalny wpływ nagrania na wizerunek jest większy, niż w przypadku wpływu na odbiór mniej znanych postaci. Ponadto jeżeli porównamy ze sobą wartości mediany odpowiedzi dla filmu pierwszego, przesegregowanej przy pomocy zmiennej E z pozostałymi filmami, zauważyć można, iż w grupie „nie rozpoznałem" dla tego filmu mediana przyjmuje wartość 6. Wiedząc, iż różnice w grupie dla pierwszego z filmów są istotne statystycznie (test Kruskal-Wallis $p < 0,05$) stwierdzić można, iż respondenci którzy nie rozpoznali deepfake uważają wpływ nagrania na odbiór osób w nich występujących, jako istotnie większy niż przez grupę osób które oszustwo rozpoznały.

6.2 Aprobata fałszywych nagrań

Elementem niezwykle istotnym w niniejszym badaniu było sprawdzenie w jaki sposób zachowają się respondenci po obejrzeniu każdego z nagrań. Poprzez mechanizmy oceniające, będące w powszechnym użyciu w każdym medium społecznościowym, internauta może w prosty sposób przyczynić się do proliferacji nieprawdziwych nagrań, zwiększając ich zasięgi oraz zatwierdzając je swoim autorytetem. Na zarówno na portalu Facebook jak i YouTube popularne są „lajki", na portalu społecznościowym X (wcześniej Twitter) zaś „serduszka". LinkedIn pozwala „polecić" daną treść, umożliwiając uszczegółowienie „polecenia" jako „gratulacje", „wsparcie", „super", „wnikliwe" oraz „śmieszne". Również inne społeczności skupione na platformach takich jak Reddit czy Discord, powszechnie używają zwrotów takich jak „plusowanie", „lajkowanie", „polubienia".

Celem sprawdzenia prawdopodobieństwa pozytywnej reakcji respondentów na nagrania, postanowiono zadać następujące pytanie: „Czy polubił/a byś ten film?". Odpowiedzi były możliwe na 10 stopniowej skali, gdzie 1 to „w ogóle", zaś 10 „bardzo".

Założono, iż respondenci w najmniejszym stopniu będą skorzy „lajkować" nagrania nieznanych osób (wideo 2 i 3). Nieco lepiej powinny zostać ocenione nagrania średnio znanych influencerów (film 4 i 5) oraz nagranie szóste (deepfake). Różnica powinna być tutaj widoczna zwłaszcza w grupie, która rozpoznała oszustwo. Najwięcej „polubień" zebrać powinien film pierwszy (deepfake). Osoba na nim prezentowana cieszy się największą popularnością, a jej nagranie zostało wykonane w najdokładniejszy sposób. Poniżej zaprezentowano statystyki opisowe do czternastego pytania.

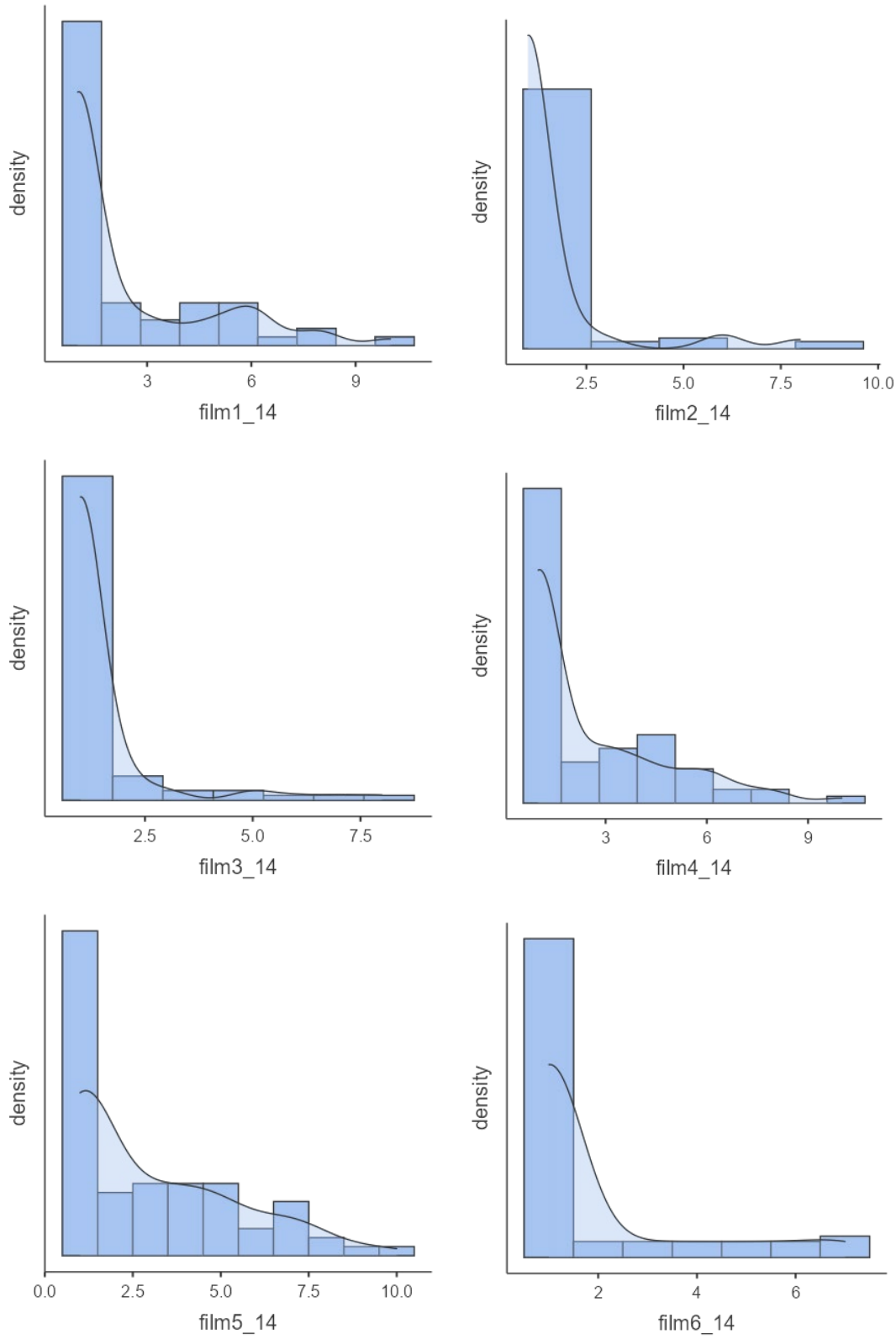
Statystyki opisowe

	film1_14	film2_14	film3_14	film4_14	film5_14	film6_14
N	60	80	79	80	80	79
Brakujące odpowiedzi	22	2	3	2	2	3
M	2.38	1.48	1.44	2.45	3.01	1.87
SE	0.295	0.163	0.150	0.240	0.267	0.201
95% CI dolna granica przedziału ufności dla średniej	1.81	1.16	1.15	1.98	2.49	1.48
95% CI górna granica przedziału ufności dla średniej	2.96	1.79	1.74	2.92	3.53	2.27
Me	1.00	1.00	1.00	1.00	2.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	2.29	1.46	1.34	2.15	2.38	1.79
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	10.0	8.00	8.00	10.0	10.0	7.00
SKE	1.61	3.49	3.52	1.49	0.991	1.93
SEk	0.309	0.269	0.271	0.269	0.269	0.271
K	1.63	11.7	12.3	1.52	0.842	2.36
Std. error K	0.608	0.532	0.535	0.532	0.532	0.535
S-W	0.668	0.369	0.382	0.725	0.817	0.548
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 53 Statystyki opisowe dla czternastego pytania „Czy polubił/a byś ten film?”.

Dane z powyższej tabeli jednoznacznie wskazują, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Wartości testu Shapiro-Wilk dla wszystkich filmów wynoszą $p < 0,001$, co sugeruje, że odrzucić należy hipotezę o normalności rozkładu. Ponadto, wartości skośności dla każdego filmu są większe od wartości błędu standardowego skośności, a w większości przypadków przekraczają 1 (z wyjątkiem filmu piątego). Dodatkowo, wartości kurtozy w każdym przypadku są wyższe od wartości błędu standardowego, co wskazuje na znaczne spłaszczenie lub spiczastość rozkładów.

Z uwagi na stwierdzony brak normalności, w kolejnych analizach statystycznych zastosowano testy nieparametryczne, które są bardziej odpowiednie w takich przypadkach. Testy te pozwalają na wiarygodną ocenę różnic między grupami bez konieczności zakładania normalności rozkładu wyników.



Wykres 56 Zbiór 6 wykresów odpowiedzi na czternaste pytanie dla każdego z sześciu filmów.

Na powyższych wykresach zauważamy wyraźną przewagę niskich odpowiedzi, co wskazuje na silną prawoskośność rozkładów. Szczególnie dominująca jest odpowiedź „1 – wcale”, co sugeruje, że respondenci są zdecydowani co do braku pozytywnego

odbioru tych nagrań. Filmy 2 i 3, które prezentowały nieznane osoby i były autentyczne, również uzyskały niskie oceny, co może sugerować, że widzowie są sceptyczni wobec treści prezentowanych przez obce twarze. Z kolei film 6, będący deepfake'em, również spotkał się z niskimi ocenami, co może sugerować, że technologia deepfake nie tylko wpływa negatywnie na odbiór filmu, ale także nie przyciąga uwagi widzów w sposób, który skłaniałby ich do ”polubienia” treści. Filmy 1 (deepfake) oraz 4 i 5 (średnio znani influencerzy) są oceniane w zbliżony sposób, jednak większość respondentów wciąż nie wykazuje chęci „polubienia” tych nagrań. To może wskazywać na ogólną nieufność wobec treści wideo, niezależnie od ich autentyczności lub popularności prezentujących je osób.

Aby zbadać, czy odpowiedzi na pytanie dotyczące chęci „polubienia” nagrania różniły się w zależności od tego, czy respondenci rozpoznali nagrania deepfake, przeprowadzono analizę z uwzględnieniem zmiennej nominalnej E: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”. Celem tej analizy było zrozumienie czy osoby, które potrafiły zidentyfikować fałsz w nagraniach, były mniej skłonne do ”polubienia” filmów, które okazały się deepfake'ami.

Statystyki opisowe

	zmienna nominalna E	**film1_14**	**film2_14**	**film3_14**	**film4_14**	**film5_14**	**film6_14**
N	nie	29	32	32	32	32	32
	tak	18	25	25	25	25	25
Brakujące odpowiedzi	nie	3	0	0	0	0	0
	tak	7	0	0	0	0	0
M	nie	3.31	1.38	1.09	2.59	2.66	2.69
	tak	1.44	1.76	1.60	2.24	2.80	1.20
SE	nie	0.509	0.228	0.0524	0.373	0.355	0.395
	tak	0.336	0.384	0.337	0.425	0.503	0.129
95% CI dolna granica przedziału ufności dla średniej	nie	2.31	0.928	0.991	1.86	1.96	1.91
	tak	0.787	1.01	0.940	1.41	1.81	0.947

Statystyki opisowe

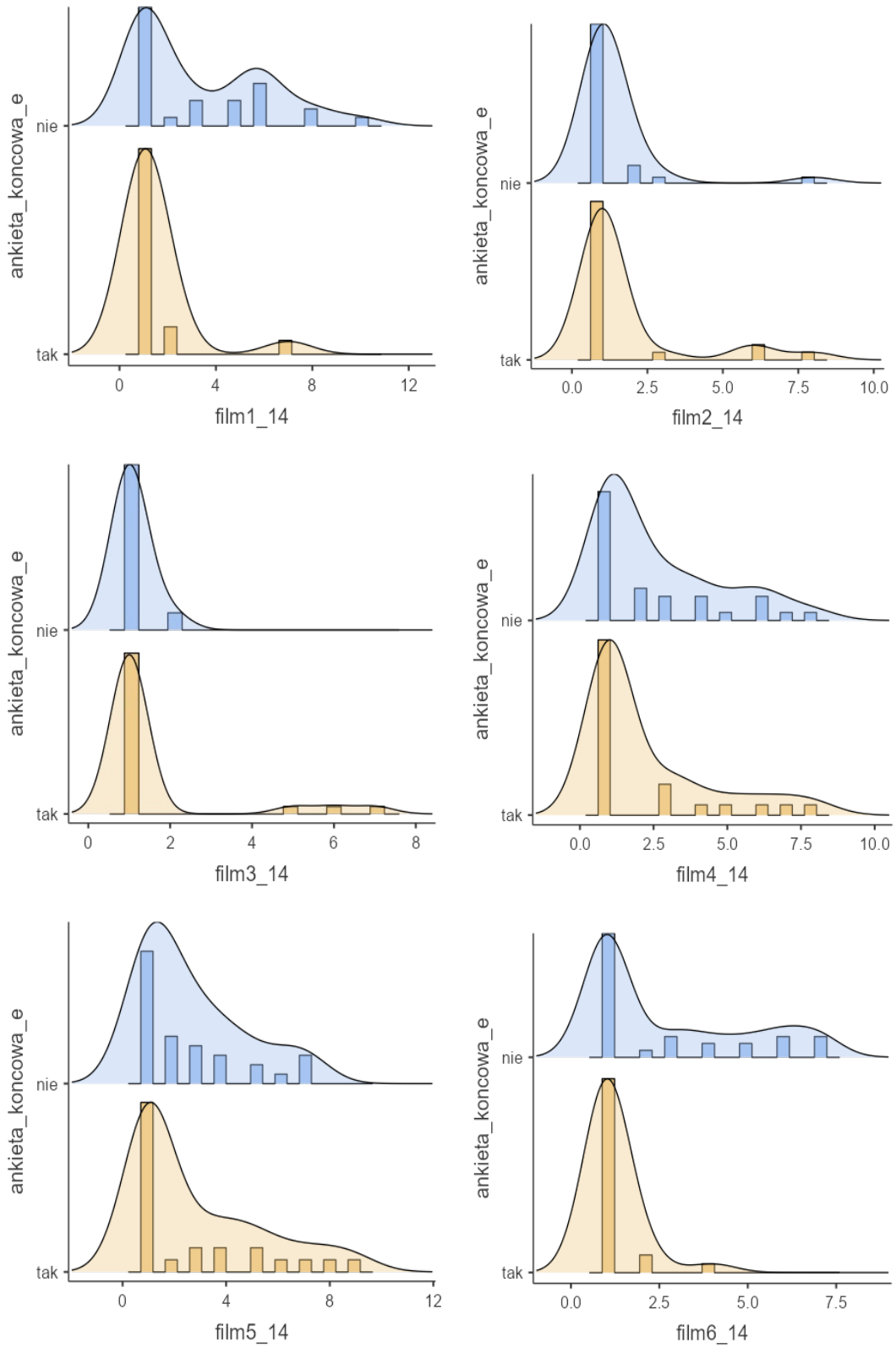
	zmienna nominalna E	film1_14	film2_14	film3_14	film4_14	film5_14	film6_14
95% CI górna granica przedziału ufności dla średniej	nie	4.31	1.82	1.20	3.32	3.35	3.46
	tak	2.10	2.51	2.26	3.07	3.79	1.45
Me	nie	2.00	1.00	1.00	1.50	2.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	2.74	1.29	0.296	2.11	2.01	2.24
	tak	1.42	1.92	1.68	2.13	2.52	0.645
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	10.0	8.00	2.00	8.00	7.00	7.00
	tak	7.00	8.00	7.00	8.00	9.00	4.00
SKE	nie	0.825	4.73	2.93	1.17	1.07	0.900
	tak	3.91	2.48	2.63	1.64	1.22	3.84
SEk	nie	0.434	0.414	0.414	0.414	0.414	0.414
	tak	0.536	0.464	0.464	0.464	0.464	0.464
K	nie	-0.478	24.1	7.00	0.226	-0.273	-0.791
	tak	15.9	5.12	5.66	1.62	0.352	15.8
Std. error K	nie	0.845	0.809	0.809	0.809	0.809	0.809
	tak	1.04	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.804	0.327	0.334	0.773	0.797	0.741
	tak	0.358	0.455	0.401	0.652	0.751	0.357
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 54 Statystyki opisowe dla czternastego pytania „Czy polubił/a byś ten film?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Tabela powyżej przedstawia dane analogiczne do wcześniejszej, z dodatkowym uwzględnieniem zmiennej nominalnej E. Wartości skośności dla wszystkich rozkładów wyników są większe niż błąd standardowy tej miary, co sugeruje, że rozkłady są asymetryczne i różnią się od normalnego. Również wartości kurtozy w większości przypadków przekraczają błąd standardowy, z wyjątkiem filmu 5 oraz odpowiedzi „nie"

dla filmów 1, 4 i 6. Niskie wartości p ($p < 0,001$) w teście Shapiro-Wilk jednoznacznie potwierdzają brak normalności rozkładów.

Poniżej znajdują się histogramy dla każdego filmu, które prezentują odpowiedzi respondentów, podzielone według zmiennej nominalnej E: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.



Wykres 57 Zbiór 6 wykresów odpowiedzi na czternaste pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Tabela oraz przedstawione powyżej histogramy ukazują różnice w odpowiedziach na czternaste pytanie pomiędzy osobami, które uznały, że rozpoznały

deepfake, a tymi, które stwierdziły, że im się to nie udało. Zgodnie z przewidywaniami, zwłaszcza w przypadku filmów pierwszego i szóstego (deepfake), w grupie poprawnie identyfikującej fałszywe nagrania przeważają odpowiedzi o niższych ocenach. Aby ocenić, czy zaobserwowane różnice między grupami są statystycznie istotne, przeprowadzono serię testów Kruskala-Wallisa. Wyniki tych analiz zostały przedstawione w tabeli poniżej.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_14	6.4001	1	0.011	0.13913
film2_14	0.0361	1	0.849	6.45e-4
film3_14	0.2052	1	0.651	0.00366
film4_14	1.0823	1	0.298	0.01933
film5_14	0.0902	1	0.764	0.00161
film6_14	7.9756	1	0.005	0.14242

Tabela 55 Test Kruskal-Wallis dla odpowiedzi do pytania czternastego.

Na podstawie przedstawionej tabeli można stwierdzić, że respondenci udzielali statystycznie zróżnicowanych odpowiedzi na pytanie w odniesieniu do filmów deepfake 1 i 6 ($p<0,05$). Zauważone różnice pomiędzy obiema grupami są na tyle znaczące, że mogą być wiarygodnie wytłumaczone za pomocą analiz statystycznych. Aby zweryfikować hipotezę zakładającą, że odpowiedzi na pytanie czternaste („Czy polubił/a byś ten film?") będą się różniły w zależności od rodzaju filmu, przeprowadzono test Friedmana. Wyniki tego testu wskazały, że odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych) istotnie się od siebie różnią (χ^2 (5) = 36,6; $p < 0,001$).

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na czternaste pytanie, przeprowadzono test porównania parami Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_14	-	film2_14	3.143	0.002
film1_14	-	film3_14	3.693	< .001
film1_14	-	film4_14	0.864	0.388
film1_14	-	film5_14	1.061	0.290
film1_14	-	film6_14	0.982	0.327
film2_14	-	film3_14	0.550	0.583

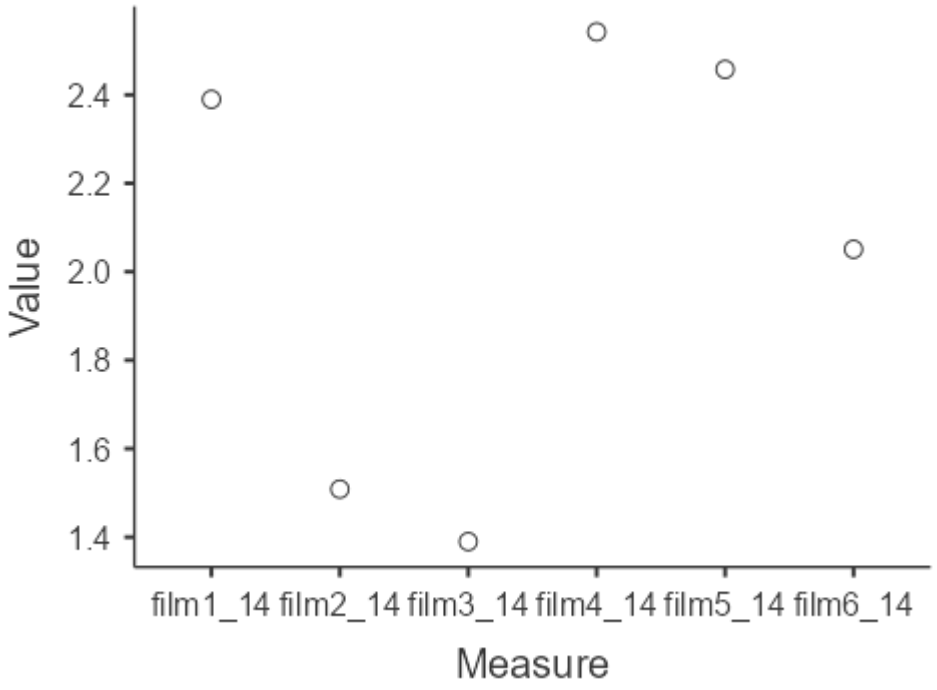
Porównania Parami (Durbin-Conover)

			Statistic	p
film2_14	-	film4_14	4.007	< .001
film2_14	-	film5_14	4.204	< .001
film2_14	-	film6_14	2.161	0.032
film3_14	-	film4_14	4.557	< .001
film3_14	-	film5_14	4.754	< .001
film3_14	-	film6_14	2.711	0.007
film4_14	-	film5_14	0.196	0.844
film4_14	-	film6_14	1.847	0.066
film5_14	-	film6_14	2.043	0.042

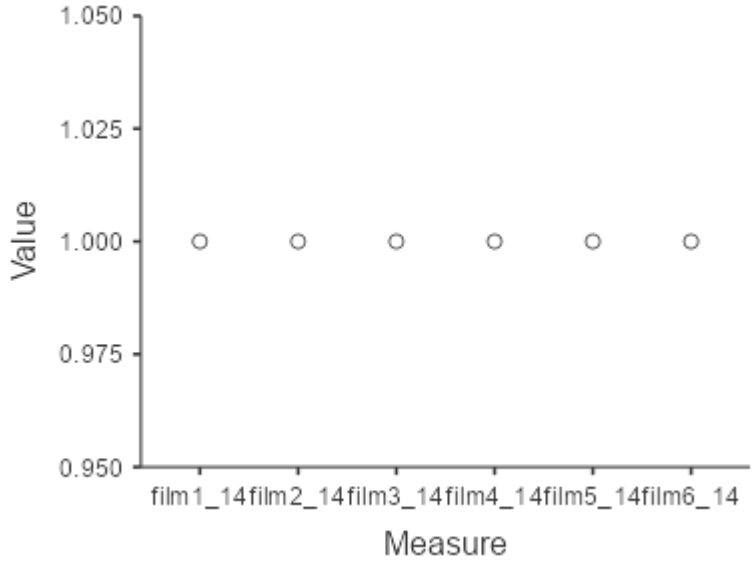
Tabela 56 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania czternastego.

Obliczenia te wykazały, że film pierwszy (deepfake) istotnie różni się od filmów 2 i 3 ($p < 0,05$), natomiast nie zaobserwowano różnic między nim, a filmami 4, 5 i 6 ($p > 0,05$). Różnic nie wykryto również między nagraniami 2 i 3 (oba prezentowały nieznane osoby). Istotne różnice pojawiły się jednak między nimi a nagraniami 4 i 5 ($p < 0,001$) oraz 6 (deepfake, $p < 0,05$). Brak różnic odnotowano między filmami 4 i 5, natomiast test Durbin-Conover wskazał na różnice między filmem 5 a 6 ($p < 0,05$).

Można zatem wnioskować, że nagrania deepfake, szczególnie film 1, byłyby „polubiane" równie chętnie, co nagrania przedstawiające średnio znanych influencerów (filmy 4 i 5). Aby wzmocnić weryfikację tej hipotezy, poniżej zamieszczono wykresy średnich i median. Uwzględniono to, że rozkłady wyników znacznie odbiegają od rozkładu normalnego, co sprawia, że średnie są podatne na zakłócenia ze strony miar tendencji centralnej.



Wykres 58 Średnia odpowiedzi dla pytania czternastego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 59 Mediana odpowiedzi dla pytania czternastego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Obserwując rozkład median na wykresie, zauważyć można iż środkowe wartości dla każdego z filmów wynoszą najmniejszą możliwą wartość, czyli 1. Jeżeli jednak porównamy ze sobą wartości mediany odpowiedzi przesegregowanej przy pomocy zmiennej E, zauważyć można, iż w grupie „nie", dla filmów 1 oraz 5 przyjmuje ona wartość 2. Wiedząc, iż różnice w grupie dla pierwszego z filmów są istotne statystycznie

(test Kruskal-Wallis $p < 0,05$) stwierdzić można, iż respondenci którzy nie rozpoznali deepfake są istotnie chętniejsi do ”polubienia” filmu deepfake, niż osoby które rozpoznały fałsz. Dzięki powyższej analizie zauważyć można, iż osoby, które nie rozpoznały deepfake statystycznie różnie oceniają swoje prawdopodobieństwo „polubienia” nagrania od przeciwnej grupy w przypadku filmów deepfake.

Kolejnym wnioskiem jest fakt, iż respondenci niechętnie „lajkują” materiały wideo prezentujące nieznane sobie osoby. Większa statystycznie szansa, iż ktoś „polubi” dany film jest wówczas, gdy prezentuje on rozpoznawalną twarz popularnej osoby zmienioną przy pomocy deepfake niż nieznanej osoby.

6.3 Propagacja fałszywych nagrań

Oprócz możliwości „lajkowania” filmów, najpopularniejsze portale społecznościowe oferują jeszcze jedną możliwość zwiększenia proliferacji danej treści. Portal Facebook umożliwia kliknięcie „udostępnij”, a następnie zamieszczenie danej treści w ”aktualnościach”, „swojej relacji”, wiadomości prywatnej, „na stronie”, „na grupie” oraz „w profilu znajomego”. Portal X (dawniej Twitter) pozwala „podać dalej wpis” umożliwiając fakultatywne uzupełnienie go treścią („cytuj”). LinkedIn przy prezentowanych postach również pozostawia możliwość dodania swojej treści („udostępnij z uwagami”) lub udostępnienie „natychmiast” na tym portalu.

Celem sprawdzenia prawdopodobieństwa „udostępnienia” przez respondentów danego nagrania, postanowiono zadać następujące pytanie: „Czy udostępnił/a byś ten film swoim znajomym?”. Odpowiedzi były możliwe na 10 stopniowej skali, gdzie 1 to „w ogóle”, zaś 10 „bardzo”.

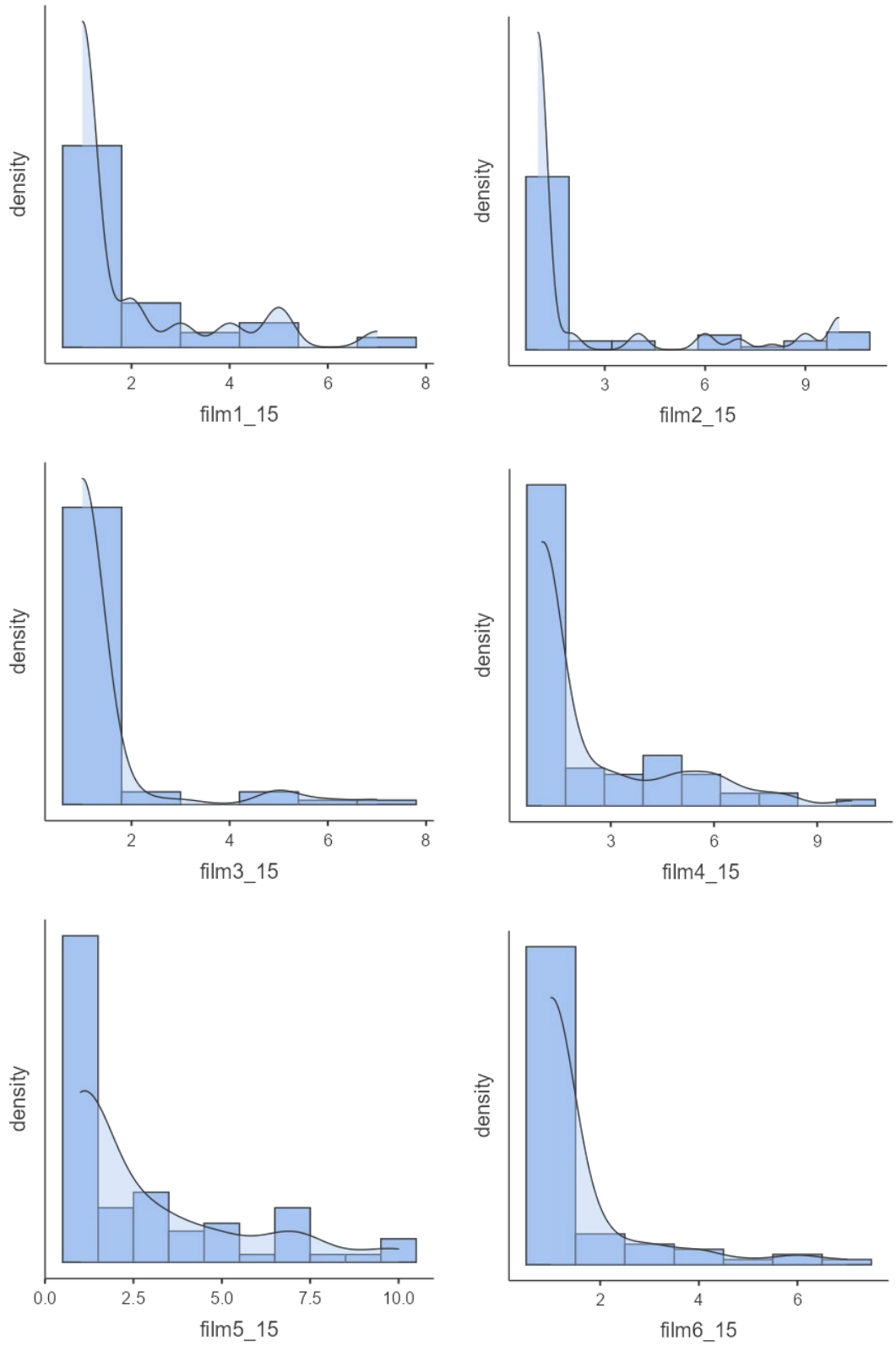
Założono, iż podobnie jak we wcześniejszym pytaniu, respondenci w najmniejszym stopniu będą skorzy „udostępnić” nagrania nieznanych osób (wideo 2 i 3). Nieco chętniej powinny zostać „udostępnione” nagrania średnio znanych influencerów (film 4 i 5) oraz nagranie szóste (deepfake). Różnica odpowiedzi na to pytanie powinna być widoczna zwłaszcza w grupie, która rozpoznała oszustwo. Najchętniej „udostępnianym” wideo powinien być film pierwszy (deepfake). Osoba na nim prezentowana cieszy się największą popularnością, a jej nagranie zostało wykonane w najdokładniejszy sposób. Poniżej zaprezentowano statystyki opisowe do piętnastego pytania.

Statystyki opisowe

	film1_15	film2_15	film3_15	film4_15	film5_15	film6_15
N	60	80	79	80	80	79
Brakujące odpowiedzi	22	2	3	2	2	3
M	1.88	2.55	1.34	2.31	2.83	1.54
SE	0.206	0.336	0.131	0.242	0.288	0.144
95% CI dolna granica przedziału ufności dla średniej	1.48	1.89	1.08	1.84	2.26	1.26
95% CI górna granica przedziału ufności dla średniej	2.29	3.21	1.60	2.79	3.39	1.83
Me	1.00	1.00	1.00	1.00	1.00	1.00
D	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.60	3.00	1.16	2.17	2.58	1.28
Min	1.00	1.00	1.00	1.00	1.00	1.00
Max	7.00	10.0	7.00	10.0	10.0	7.00
SKE	1.83	1.70	3.59	1.63	1.38	2.69
SEk	0.309	0.269	0.271	0.269	0.269	0.271
K	2.44	1.28	12.3	1.78	0.876	6.97
Std. error K	0.608	0.532	0.535	0.532	0.532	0.535
S-W	0.620	0.562	0.328	0.670	0.738	0.497
PS-W	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 57 Statystyki opisowe dla piętnastego pytania „Czy udostępnił/a byś ten film swoim znajomym?”.

Dane z powyższej tabeli wyraźnie wskazują, że rozkład uzyskanych wyników nie jest zbliżony do rozkładu normalnego. Potwierdzają to wyniki testu Shapiro-Wilk, w którym za każdym razem uzyskano wartość $p < 0{,}001$, oraz wartości skośności, które dla każdego filmu przewyższają wartość błędu standardowego tej miary i we wszystkich przypadkach przekraczają 1. Dodatkowo, brak normalnego rozkładu potwierdzają wartości kurtozy, które również we wszystkich przypadkach są wyższe niż błąd standardowy tej miary. W związku z brakiem normalności rozkładu wyników, w kolejnych analizach statystycznych zastosowano testy nieparametryczne.



Wykres 60 Zbiór 6 wykresów odpowiedzi na piętnaste pytanie dla każdego z sześciu filmów.

Podobnie jak w poprzednim podrozdziale, wykresy przedstawiają wyraźną przewagę niskich ocen. Uzyskane dane wskazują na silną prawoskośność rozkładów, z dominacją odpowiedzi „1 – wcale”, zwłaszcza w przypadku filmów 2, 3 i 6. Tylko film

6 był deepfake'iem, natomiast filmy 2 i 3 to prawdziwe nagrania prezentujące nieznane osoby. Film 1 (deepfake) oraz filmy 4 i 5 (średnio znani influencerzy) zostały ocenione w sposób zbliżony, choć większość respondentów nadal deklarowała brak chęci „udostępnienia" tych nagrań („w ogóle") swoim znajomym.

Aby zweryfikować, czy odpowiedzi na pytanie dotyczące chęci „udostępnienia" nagrania znajomym różniły się w zależności od tego, czy respondent rozpoznał deepfake, podjęto dalszą analizę z uwzględnieniem zmiennej nominalnej E. Zmienna ta brzmiała: „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?". Kluczowe było sprawdzenie czy osoby, które rozpoznały fałszywe nagrania, były mniej skłonne do "udostępnienia" tych filmów niż osoby, którym się to nie udało.

Statystyki opisowe

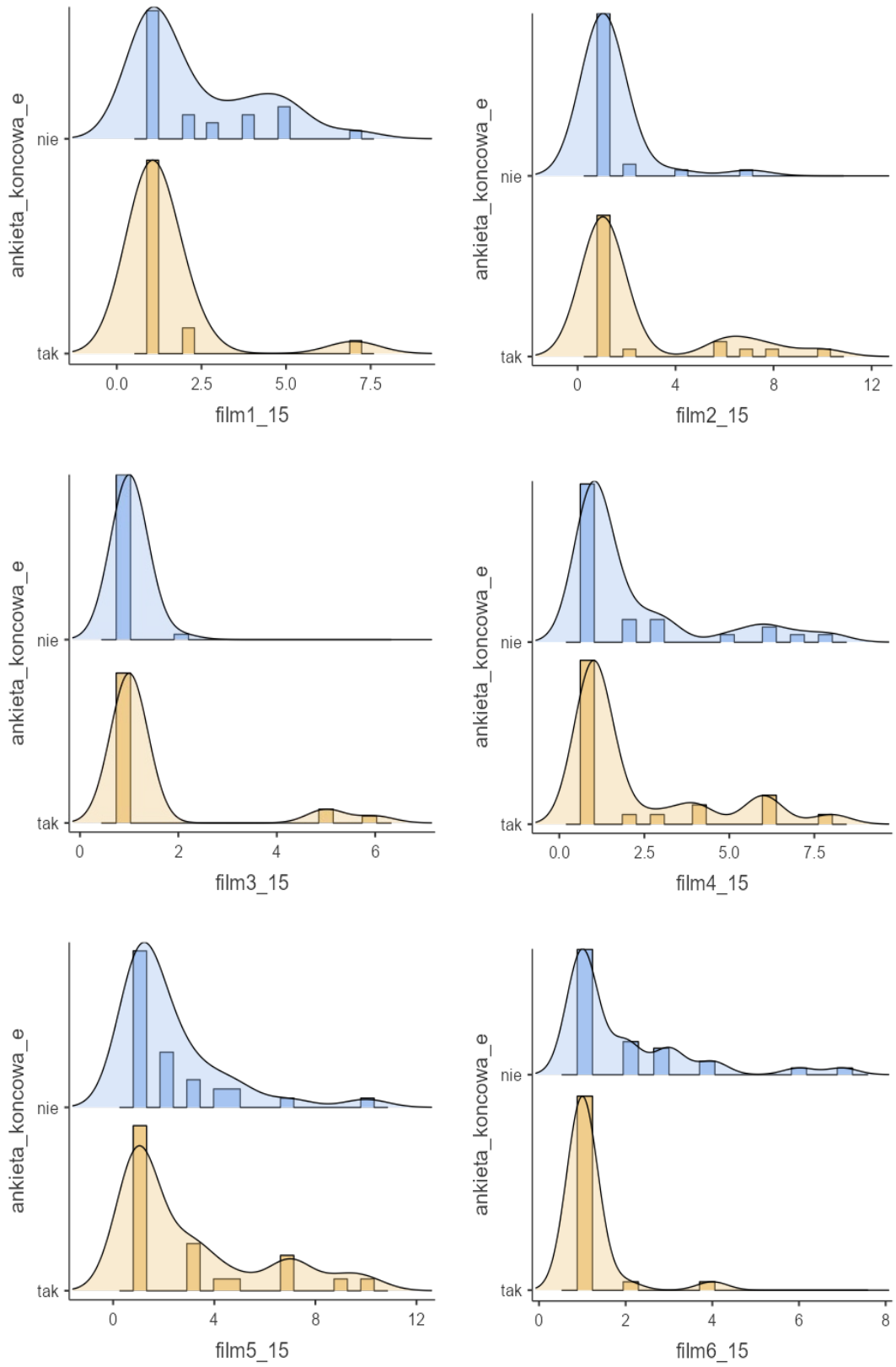
	zmienna nominalna E	**film1_15**	**film2_15**	**film3_15**	**film4_15**	**film5_15**	**film6_15**
N	nie	29	32	32	32	32	32
	tak	18	25	25	25	25	25
Brakujące odpowiedzi	nie	3	0	0	0	0	0
	tak	7	0	0	0	0	0
M	nie	2.31	1.34	1.03	2.13	2.28	1.94
	tak	1.44	2.32	1.52	2.24	3.00	1.16
SE	nie	0.330	0.209	0.0313	0.356	0.365	0.269
	tak	0.336	0.538	0.289	0.425	0.569	0.125
95% CI dolna granica przedziału ufności dla średniej	nie	1.66	0.935	0.970	1.43	1.57	1.41
	tak	0.787	1.27	0.953	1.41	1.89	0.915
95% CI górna granica przedziału ufności dla średniej	nie	2.96	1.75	1.09	2.82	3.00	2.47
	tak	2.10	3.37	2.09	3.07	4.11	1.40
Me	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
D	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
SD	nie	1.77	1.18	0.177	2.01	2.07	1.52

	zmienna nominalna E	film1_15	film2_15	film3_15	film4_15	film5_15	film6_15
	tak	1.42	2.69	1.45	2.13	2.84	0.624
Min	nie	1.00	1.00	1.00	1.00	1.00	1.00
	tak	1.00	1.00	1.00	1.00	1.00	1.00
Max	nie	7.00	7.00	2.00	8.00	10.0	7.00
	tak	7.00	10.0	6.00	8.00	10.0	4.00
SKE	nie	1.10	4.16	5.66	1.82	2.26	1.98
	tak	3.91	1.85	2.56	1.55	1.28	4.35
SEk	nie	0.434	0.414	0.414	0.414	0.414	0.414
	tak	0.536	0.464	0.464	0.464	0.464	0.464
K	nie	0.620	18.3	32.0	2.21	5.73	3.86
	tak	15.9	2.14	5.11	1.19	0.440	19.7
Std. error K	nie	0.845	0.809	0.809	0.809	0.809	0.809
	tak	1.04	0.902	0.902	0.902	0.902	0.902
S-W	nie	0.752	0.334	0.172	0.629	0.680	0.678
	tak	0.358	0.556	0.398	0.645	0.737	0.286
PS-W	nie	< .001	< .001	< .001	< .001	< .001	< .001
	tak	< .001	< .001	< .001	< .001	< .001	< .001

Tabela 58 Statystyki opisowe dla piętnastego pytania „Czy udostępnił/a byś ten film swoim znajomym?" z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?".

Powyższa tabela przedstawia analogiczne dane jak wcześniejszy diagram, z uwzględnieniem wspomnianej zmiennej nominalnej E. Podobnie jak we wcześniejszej tabeli, rozkłady wyników we wszystkich analizowanych przypadkach odbiegają od rozkładu normalnego. Potwierdzają to wartości skośności, które we wszystkich przypadkach przewyższają błąd standardowy tej miary. Kolejnym wskaźnikiem braku normalności są wartości kurtozy — w większości przypadków wartości bezwzględne tej miary przewyższają błąd standardowy (wyjątkiem jest film pierwszy w grupie „nie" oraz nagranie piąte w grupie „tak"). Niskie wyniki testu Shapiro-Wilk, we wszystkich przypadkach wynoszące $p < 0{,}001$, również wskazują na brak normalnego rozkładu.

Poniżej przedstawiono histogramy dla każdego filmu, z podziałem respondentów na grupy według zmiennej nominalnej E („Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?"), co pozwala zobaczyć rozkład odpowiedzi w zależności od tego, czy badani uważają, że rozpoznali deepfake.



Wykres 61 Zbiór 6 wykresów odpowiedzi na piętnaste pytanie dla każdego z sześciu filmów z uwzględnieniem zmiennej nominalnej „Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake?”.

Powyższa tabela oraz histogramy przedstawiają różnice w odpowiedziach na pytanie piętnaste, w zależności od tego, czy respondenci uważali, że rozpoznali deepfake, czy nie. Zgodnie z oczekiwaniami, w szczególności dla filmów 1 i 6 (oba deepfake), w grupie, która rozpoznała fałszywe nagrania, przeważały niższe oceny. Aby ocenić, czy różnice między grupami są statystycznie istotne, przeprowadzono serię testów Kruskala-Wallisa. Wyniki tych testów zostały przedstawione w poniższej tabeli.

Kruskal-Wallis

	χ^2	df	p	ε^2
film1_15	4.045	1	0.044	0.08793
film2_15	1.623	1	0.203	0.02898
film3_15	1.805	1	0.179	0.03224
film4_15	3.68e-4	1	0.985	6.57e-6
film5_15	0.226	1	0.635	0.00403
film6_15	7.416	1	0.006	0.13243

Tabela 59 Test Kruskal-Wallis dla odpowiedzi do pytania piętnastego.

Na podstawie powyższej tabeli wnioskować można, iż respondenci w statystycznie różny sposób odpowiadali na to pytanie w przypadku filmów 1 i 6 (deepfake) ($p<0,05$). Podobnie jak we wcześniejszym podrozdziale, stwierdzone różnice pomiędzy tymi dwoma grupami są na tyle duże, aby móc je wiarygodnie wyjaśnić przy użyciu statystyk.

W celu testowania hipotezy mówiącej o tym, że odpowiedź na pytanie piętnaste („Czy udostępnił/a byś ten film swoim znajomym?") będzie się różniła w przypadku różnych filmów, przeprowadzono test Friedmana. Test ten wykazał, iż odpowiedzi na to pytanie w przypadku wszystkich filmów (zarówno prawdziwych, jak i fałszywych), różnią się między sobą (χ^2 (5) = 44,5; $p < 0,001$).

Celem zbadania różnic pomiędzy poszczególnymi odpowiedziami na drugie pytanie, przeprowadzono test Durbin-Conover. Tabela z wynikami testu znajduje się poniżej.

Porównania Parami (Durbin-Conover)

			Statistic	p
film1_15	-	film2_15	4.392	< .001
film1_15	-	film3_15	4.568	< .001
film1_15	-	film4_15	1.142	0.254
film1_15	-	film5_15	0.922	0.357
film1_15	-	film6_15	2.416	0.016

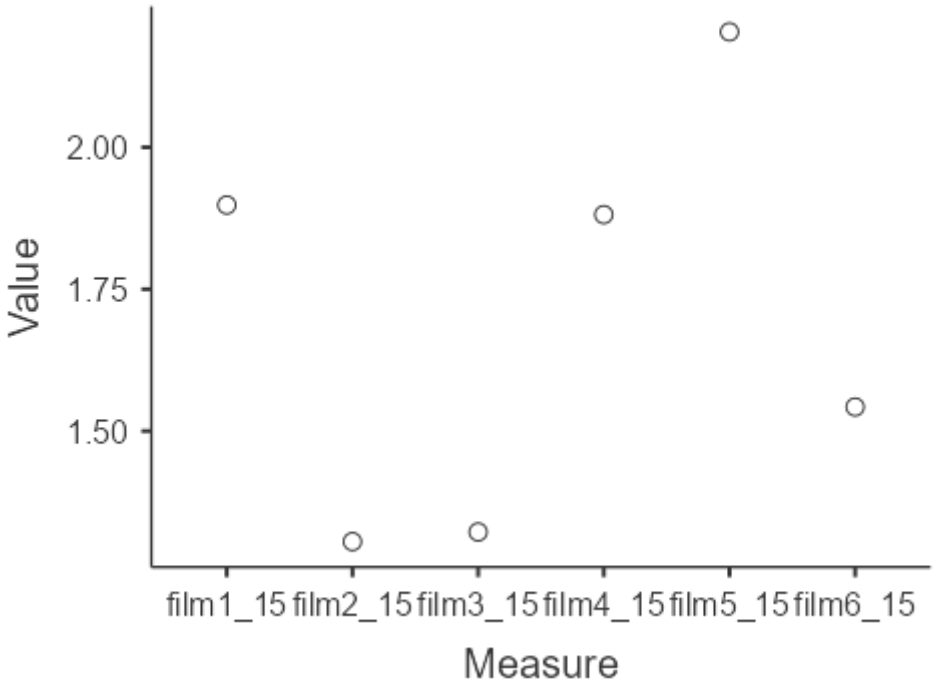
Porównania Parami (Durbin-Conover)

			Statistic	p
film2_15	-	film3_15	0.176	0.861
film2_15	-	film4_15	3.250	0.001
film2_15	-	film5_15	5.315	< .001
film2_15	-	film6_15	1.977	0.049
film3_15	-	film4_15	3.426	< .001
film3_15	-	film5_15	5.490	< .001
film3_15	-	film6_15	2.152	0.032
film4_15	-	film5_15	2.064	0.040
film4_15	-	film6_15	1.274	0.204
film5_15	-	film6_15	3.338	< .001

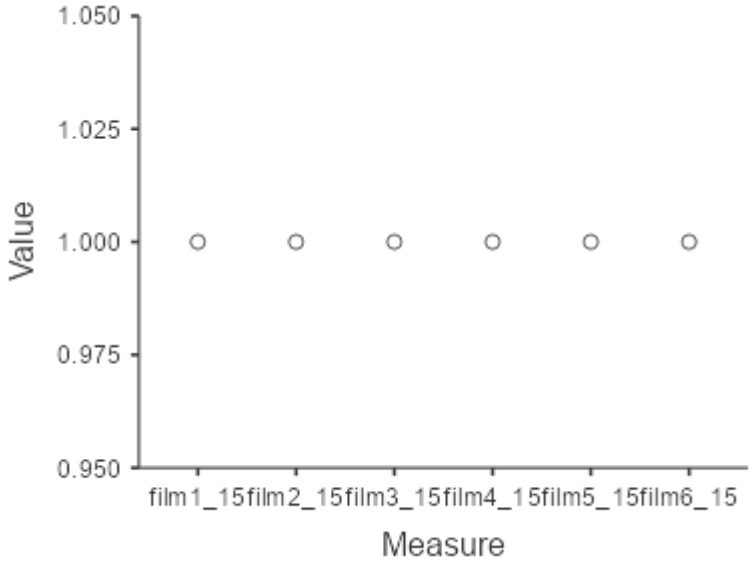
Tabela 60 Test porównywania parami Durbin-Conover dla odpowiedzi do pytania piętnastego.

Powyższe obliczenia pokazały, że film 1 (deepfake) różni się od filmu 2 i 3 ($p < 0,001$) oraz drugiego z filmów deepfake, filmu 6 ($p < 0,05$), Nie zaobserwowano różnicy pomiędzy filmem 1, a filmami 4 oraz 5 ($p > 0,05$). Pomiędzy nagraniami 2, a 3 również nie zaobserwowano różnic (oba prezentowały nieznane osoby). Różnica jest natomiast między nimi, a nagraniami 4, 5 ($p < 0,001$) oraz 6 (deepfake, $p < 0,05$). Nie zaobserwowano różnic między filmami 4, a 5, natomiast z porównania Durbin-Conover wiemy, że różnice występują między filmem 5, a 6 ($p < 0,001$).

Można wnioskować, że nagrania deepfake, zwłaszcza pierwsze, byłyby „udostępniane" znajomym równie chętnie jak nagrania średnio znanych influencerów (filmy 4 i 5). Aby zweryfikować tę hipotezę, poniżej przedstawiono wykresy średnich i median. Ze względu na to, że rozkłady wyników znacznie odbiegają od rozkładu normalnego, średnie są w tym przypadku podatne na zniekształcenie przez miary tendencji centralnej.



Wykres 62 Średnia odpowiedzi dla pytania piętnastego dla poszczególnych filmów. Oś pionowa to wartości średniej, zaś pozioma odpowiadające im oznaczenia filmów.



Wykres 63 Mediana odpowiedzi dla pytania piętnastego dla poszczególnych filmów. Oś pionowa to wartości mediany, zaś pozioma odpowiadające im oznaczenia filmów.

Obserwując rozkład median na wykresie, zauważyć można, iż środkowe wartości dla każdego z filmów wynoszą najmniejszą możliwą wartość, czyli 1. Różnicę pomiędzy chęcią udostępnienia filmów zauważyć jednak można na wykresie średniej, gdzie wiedząc, iż nie zaobserwowano różnicy pomiędzy filmem 1 (deepfake), a prawdziwymi filmami średnio znanych influencerów 4 oraz 5 ($p > 0,05$), wskazać możemy, iż

respondenci „udostępniliby" je podobnie chętnie. Dzieje się tak zwłaszcza, jeżeli respondentów podzielimy przy użyciu zmiennej nominalnej E. Wówczas średnia dla filmów deepfake wynosi odpowiednio 2.31 dla filmu pierwszego oraz 1.94 dla nagrania ostatniego. Wiedząc, iż różnice w grupach dla obu z filmów są istotne statystycznie (test Kruskal-Wallis $p < 0,05$) stwierdzić można, iż respondenci, którzy nie rozpoznali deepfake są istotnie chętniejsi do przekazania dalej filmu deepfake, niż osoby, które rozpoznały fałsz.

Kolejnym wnioskiem jest fakt, iż respondenci niechętnie „udostępniają" materiały wideo prezentujące nieznane sobie osoby. Większa statystycznie szansa, iż ktoś „udostępni" swoim znajomym dany film jest wówczas, gdy prezentuje on rozpoznawalną twarz lub twarz popularnej osoby zmienioną przy pomocy deepfake.

6.4 Analiza korelacji moderatorów

Niniejszy podrozdział ma na celu wykrycie cech osobowych, które mogły mieć wpływ na udzielane odpowiedzi, a które powodować mogą na przykład większą skłonność do nierozpoznawania fałszywych filmów lub zwiększoną podatność na ich przekaz.

Na potrzeby niniejszego badania postanowiono sprawdzić czy odpowiedź na poszczególne pytania różnić się może w zależności od personalnych cech osobowych. W poniższym podrozdziale omówione zostały korelacje moderatorów psychologicznych dla poszczególnych odpowiedzi na pytania ankietowe.

W związku z brakiem rozkładu normalnego zmiennych, przeprowadzono serię nieparametrycznych korelacji Rho Spearmana. Jest ono znane jest również jako korelacja rang Spearmana lub po prostu rho (ρ). Jest to jedna z nieparametrycznych miar monotonicznej zależności statystycznej między zmiennymi losowymi. Miara ta używana jest do określenia stopnia związku monotonicznego pomiędzy dwiema zmiennymi. Jest ona szczególnie przydatna w przypadku, gdy dane nie spełniają założeń normalności lub gdy występują obserwacje odstające. Rho Spearmana mierzy siłę i kierunek związku pomiędzy dwiema zmiennymi, ale nie zakłada konkretnego kształtu tej zależności[267].

Współczynnik ten został opisany i rozpropagowany w 1904 roku przez angielskiego psychologa Charlesa Spearmana. Ch. Spearman zdefiniował swój

[267] D. G. Bonett, T. A. Wright, „Sample size requirements for Pearson, Kendall, and Spearman correlations", Psychometrika 65, 2000, s. 23–28.

współczynnik jako zwykły współczynnik korelacji Pearsona, liczony dla rang zmiennych (stąd nazwa współczynnik korelacji rang)[268].

Obliczanie współczynnika korelacji rang Spearmana odbywa się w dwóch etapach, najpierw wykonywane jest rangowanie, czyli zastąpienie każdej zaobserwowanej wartości przez jej numer w zbiorze posortowanym rosnąco, a następnie następuje obliczenie zwykłego współczynnika korelacji liniowej Pearsona.

Wzór na korelację rho − Spearmana prezentuje się następująco:

$\rho = 1 - 6 * \sum d^2 / (n * (n^2 - 1))$

gdzie:

ρ – to współczynnik korelacji rang Spearmana

d – to różnica rang dla danej obserwacji

n – to liczba obserwacji

Poniżej zaprezentowano 6 tabel z wartościami korelacji odpowiedzi względem zastosowanych w badaniu moderatorów psychologicznych, oddzielnie dla każdego filmu. W badaniu zastosowano 6 losowo wyświetlanych moderatorów: moderator lęku – GAD 7 – 7 pytań, moderator na depresyjność – PHQ-9 – 9 pytań, moderator potrzeby poznawczego domknięcia – 15 pytań, moderator kodów moralnych – MFQ – 30 pytań, moderator samooceny – SES – 10 pytań oraz moderator impulsywności – BIS-Brief – 8 pytań. Wszystkie moderatory zostały szczegółowo opisane w rozdziale 1.2.2.2.6.

[268] Ch. Spearman. „The proof and measurement of association between two things", Americal Journal of Psychology, 15 (1904), s. 72–101.

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swiadomosc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielokrotnosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
fim1_01	p	0.003	0.249	0.101	0.153	0.031	-0.317	-0.229	-0.011	-0.069	-0.058	0.044	-0.034	0.041	0.111	0.002
	p-value	0.982	0.026	0.372	0.174	0.782	0.004	0.041	0.922	0.547	0.613	0.701	0.768	0.718	0.329	0.987
fim1_02	p	0.012	0.138	0.129	0.229	0.076	-0.322	-0.218	-0.037	0.009	-0.051	0.070	-0.020	0.058	0.044	-0.041
	p-value	0.912	0.223	0.254	0.041	0.502	0.004	0.052	0.746	0.938	0.652	0.538	0.858	0.612	0.696	0.718
fim1_03	p	0.045	0.157	0.185	0.209	0.083	-0.299	-0.175	-0.075	0.029	-0.062	0.119	0.006	0.130	0.048	-0.042
	p-value	0.693	0.164	0.101	0.063	0.464	0.007	0.121	0.507	0.799	0.589	0.296	0.959	0.255	0.670	0.713
fim1_04	p	-0.015	0.109	0.159	0.237	0.112	-0.279	-0.248	-0.175	0.041	-0.088	-0.028	-0.057	0.135	-0.052	-0.071
	p-value	0.897	0.338	0.158	0.034	0.325	0.012	0.026	0.120	0.722	0.442	0.808	0.619	0.234	0.646	0.533
fim1_05	p	-0.046	-0.002	0.227	0.281	0.138	-0.198	-0.174	-0.159	0.148	-0.104	0.031	0.025	0.115	-0.074	-0.051
	p-value	0.689	0.987	0.043	0.011	0.224	0.078	0.123	0.160	0.194	0.361	0.789	0.830	0.315	0.517	0.655
fim1_06	p	-0.031	0.108	0.242	0.298	0.164	-0.264	-0.154	-0.099	0.081	-0.047	0.001	-0.123	0.044	-0.045	-0.111
	p-value	0.786	0.340	0.031	0.007	0.146	0.018	0.173	0.381	0.480	0.680	0.996	0.278	0.699	0.694	0.326
fim1_07	p	-0.084	0.012	-0.240	-0.137	-0.167	0.177	0.082	0.083	-0.091	0.132	0.032	0.011	-0.189	0.034	0.016
	p-value	0.459	0.917	0.032	0.224	0.139	0.117	0.471	0.464	0.426	0.248	0.779	0.926	0.095	0.767	0.885
fim1_08	p	-0.051	0.282	0.082	0.008	-0.031	-0.208	0.027	0.096	-0.055	0.033	0.180	-0.145	-0.092	0.154	0.095
	p-value	0.655	0.011	0.472	0.943	0.788	0.064	0.812	0.399	0.629	0.775	0.112	0.203	0.418	0.172	0.403
fim1_09	p	-0.055	0.303	0.033	0.095	0.006	-0.280	-0.152	0.032	0.013	0.030	0.304	-0.042	0.058	0.147	0.082
	p-value	0.629	0.006	0.773	0.403	0.958	0.012	0.178	0.781	0.912	0.793	0.006	0.713	0.613	0.194	0.469
fim1_10	p	0.016	0.059	0.113	0.109	0.050	-0.286	-0.204	-0.021	0.183	-0.054	0.144	0.020	0.196	0.116	0.074
	p-value	0.888	0.606	0.317	0.335	0.660	0.010	0.070	0.856	0.106	0.634	0.206	0.858	0.083	0.305	0.516
fim1_11	p	-0.178	0.085	-0.041	-0.023	-0.108	-0.176	-0.077	0.120	-0.011	-0.108	0.258	-0.066	0.112	0.194	0.155
	p-value	0.114	0.452	0.718	0.841	0.339	0.119	0.500	0.288	0.922	0.343	0.022	0.563	0.326	0.085	0.171
fim1_12	p	-0.186	0.276	-0.094	-0.010	-0.058	-0.290	-0.134	0.067	-0.043	0.064	0.260	-0.015	0.030	0.268	0.186
	p-value	0.099	0.013	0.406	0.929	0.607	0.009	0.235	0.554	0.704	0.578	0.021	0.897	0.793	0.016	0.098
fim1_13	p	-0.063	0.324	-0.115	-0.029	-0.112	-0.279	-0.143	0.111	-0.215	-0.033	0.259	0.050	0.138	0.130	0.248
	p-value	0.635	0.012	0.382	0.824	0.395	0.031	0.276	0.400	0.103	0.803	0.048	0.708	0.297	0.322	0.056
fim1_14	p	0.128	0.159	0.269	0.237	0.084	-0.305	-0.194	-0.104	0.085	-0.194	0.163	0.092	0.090	-0.012	0.047
	p-value	0.330	0.225	0.038	0.068	0.524	0.018	0.137	0.429	0.524	0.141	0.216	0.490	0.499	0.929	0.723
fim1_15	p	-0.013	0.110	0.178	0.218	0.041	-0.339	-0.342	-0.027	0.102	-0.183	0.225	0.101	0.080	0.046	0.072
	p-value	0.921	0.402	0.173	0.095	0.758	0.008	0.007	0.840	0.440	0.165	0.086	0.446	0.545	0.725	0.584

Tabela 61Macierz korelacji moderatorów psychologicznych dla nagrania pierwszego.

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swiadomosc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielo znacznosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
fim2_01	p	-0.214	0.042	0.008	0.139	0.009	-0.068	0.138	0.106	0.035	-0.090	-0.190	-0.238	-0.112	0.030	0.100
	p-value	0.056	0.711	0.945	0.217	0.933	0.549	0.223	0.348	0.758	0.432	0.094	0.035	0.325	0.794	0.380
fim2_02	p	-0.203	-0.021	0.252	0.029	0.112	0.089	0.250	0.071	0.043	0.081	-0.185	-0.141	-0.091	-0.091	-0.059
	p-value	0.071	0.853	0.024	0.797	0.323	0.430	0.026	0.532	0.708	0.479	0.103	0.214	0.424	0.420	0.600
fim2_03	p	-0.202	0.015	-0.028	-0.055	-0.101	-0.107	-0.016	0.129	-0.046	-0.054	-0.241	-0.144	-0.037	0.084	0.046
	p-value	0.072	0.892	0.807	0.627	0.371	0.346	0.887	0.255	0.687	0.638	0.033	0.207	0.747	0.457	0.688
fim2_04	p	-0.054	0.011	0.044	0.048	-0.038	0.072	0.096	0.132	0.063	0.030	-0.193	-0.085	-0.134	-0.120	-0.021
	p-value	0.631	0.925	0.697	0.675	0.740	0.527	0.399	0.244	0.582	0.796	0.089	0.459	0.238	0.289	0.850
fim2_05	p	-0.180	0.056	0.056	0.102	-0.025	-0.002	0.034	0.148	-0.039	-0.068	-0.282	-0.138	-0.168	-0.004	0.027
	p-value	0.109	0.620	0.623	0.369	0.824	0.986	0.764	0.190	0.734	0.553	0.012	0.226	0.139	0.973	0.813
fim2_06	p	-0.207	-0.046	0.055	0.066	-0.060	0.083	0.140	0.136	0.058	-0.012	-0.169	-0.113	-0.050	-0.038	0.007
	p-value	0.065	0.688	0.627	0.559	0.597	0.466	0.216	0.231	0.613	0.914	0.136	0.322	0.659	0.740	0.949
fim2_07	p	0.049	0.205	-0.040	0.150	0.183	0.091	-0.086	-0.065	0.008	0.097	0.123	0.111	0.047	-0.065	0.063
	p-value	0.665	0.068	0.728	0.185	0.104	0.421	0.450	0.568	0.945	0.397	0.279	0.332	0.683	0.565	0.581
fim2_08	p	-0.097	0.241	0.162	0.047	-0.107	-0.197	0.119	0.123	-0.179	0.010	0.006	-0.182	-0.004	-0.023	0.028
	p-value	0.394	0.032	0.152	0.682	0.344	0.081	0.294	0.278	0.115	0.932	0.961	0.109	0.975	0.841	0.806
fim2_09	p	-0.157	0.252	0.149	0.199	0.033	-0.106	-0.007	0.012	-0.067	-0.008	0.022	-0.156	-0.047	-0.121	0.047
	p-value	0.165	0.024	0.187	0.077	0.772	0.352	0.952	0.913	0.599	0.946	0.845	0.170	0.681	0.286	0.678
fim2_10	p	-0.195	0.050	-0.014	-0.025	-0.067	0.027	0.076	0.128	-0.003	0.071	-0.207	-0.183	0.006	-0.027	0.046
	p-value	0.083	0.659	0.902	0.824	0.554	0.809	0.500	0.258	0.982	0.536	0.067	0.107	0.956	0.814	0.685
fim2_11	p	-0.245	0.028	-0.019	-0.032	-0.130	-0.039	-0.002	0.133	-0.055	0.105	-0.245	-0.258	-0.118	0.058	-0.025
	p-value	0.028	0.806	0.869	0.779	0.250	0.729	0.988	0.238	0.631	0.356	0.029	0.022	0.301	0.610	0.824
fim2_12	p	-0.007	0.244	-0.057	0.087	0.067	-0.084	-0.104	0.040	0.017	-0.003	-0.079	-0.083	-0.036	0.056	0.095
	p-value	0.948	0.029	0.618	0.443	0.555	0.461	0.357	0.724	0.882	0.982	0.489	0.466	0.749	0.621	0.403
fim2_13	p	-0.232	-0.080	-0.106	0.045	-0.118	-0.144	-0.099	0.095	0.001	-0.052	-0.180	-0.109	-0.116	0.077	0.113
	p-value	0.039	0.480	0.350	0.689	0.297	0.201	0.380	0.400	0.990	0.646	0.112	0.338	0.308	0.496	0.319
fim2_14	p	-0.165	0.027	0.184	0.196	0.076	0.069	0.125	0.082	0.004	-0.011	-0.153	-0.083	-0.171	-0.082	-0.085
	p-value	0.143	0.813	0.102	0.081	0.503	0.546	0.269	0.472	0.970	0.925	0.178	0.466	0.131	0.467	0.453
fim2_15	p	-0.069	-0.184	0.065	-0.018	-0.075	-0.054	0.009	0.080	0.031	0.003	-0.126	-0.087	0.095	-0.035	-0.148
	p-value	0.541	0.102	0.566	0.871	0.507	0.635	0.937	0.479	0.788	0.982	0.270	0.446	0.407	0.758	0.190

Tabela 62 Macierz korelacji moderatorów psychologicznych dla nagrania drugiego.

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swiadomosc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielokrotnosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
fim3_01	p	-0.099	0.121	0.067	0.086	0.008	0.002	0.111	0.140	-0.005	0.028	0.068	-0.064	0.033	0.136	0.025
	p-value	0.388	0.290	0.559	0.450	0.942	0.988	0.329	0.218	0.968	0.811	0.554	0.579	0.772	0.233	0.830
fim3_02	p	-0.102	0.062	0.083	0.096	0.066	0.037	0.056	0.060	-0.007	0.061	-0.068	-0.099	0.104	0.109	0.037
	p-value	0.371	0.590	0.469	0.401	0.561	0.748	0.624	0.597	0.948	0.595	0.554	0.386	0.363	0.340	0.743
fim3_03	p	-0.073	0.016	0.091	0.129	0.088	-0.007	0.011	0.077	0.100	0.055	-0.109	-0.112	-0.023	-0.014	-0.019
	p-value	0.524	0.889	0.428	0.259	0.443	0.953	0.925	0.503	0.383	0.632	0.342	0.328	0.844	0.903	0.865
fim3_04	p	-0.051	0.057	0.096	0.156	0.121	0.032	0.012	0.007	0.049	-0.098	-0.161	-0.056	-0.075	-0.093	-0.068
	p-value	0.653	0.616	0.401	0.169	0.287	0.779	0.917	0.953	0.668	0.394	0.160	0.625	0.517	0.414	0.551
fim3_05	p	-0.138	0.126	0.069	0.142	0.050	-0.031	-0.024	0.021	0.025	-0.035	-0.224	-0.114	-0.181	-0.077	-0.057
	p-value	0.226	0.269	0.548	0.212	0.659	0.788	0.831	0.853	0.828	0.764	0.048	0.322	0.112	0.498	0.616
fim3_06	p	-0.173	0.063	0.010	0.118	0.013	-0.005	-0.004	0.033	0.045	0.033	-0.136	-0.068	-0.098	-0.014	-0.005
	p-value	0.127	0.578	0.930	0.301	0.910	0.963	0.970	0.775	0.693	0.774	0.236	0.555	0.395	0.900	0.962
fim3_07	p	0.071	0.075	0.049	-0.047	0.020	0.056	0.074	-0.015	0.036	0.036	0.042	-0.092	-0.126	-0.040	-0.018
	p-value	0.536	0.512	0.668	0.681	0.862	0.626	0.515	0.896	0.756	0.753	0.713	0.424	0.274	0.729	0.875
fim3_08	p	0.016	0.236	0.001	0.004	-0.046	-0.207	-0.019	0.134	-0.054	-0.233	0.137	-0.118	0.044	0.178	0.063
	p-value	0.891	0.036	0.993	0.974	0.684	0.067	0.867	0.240	0.640	0.040	0.230	0.305	0.699	0.118	0.583
fim3_09	p	0.001	0.178	0.151	0.200	-0.011	-0.074	0.042	0.000	-0.002	-0.153	0.161	-0.086	0.132	-0.010	0.006
	p-value	0.993	0.117	0.185	0.077	0.926	0.518	0.711	0.999	0.987	0.181	0.159	0.454	0.250	0.930	0.957
fim3_10	p	-0.219	0.081	-0.181	0.042	-0.112	-0.083	-0.048	0.132	-0.067	0.096	-0.155	-0.162	-0.125	0.124	0.130
	p-value	0.052	0.480	0.111	0.712	0.324	0.468	0.673	0.245	0.562	0.405	0.176	0.156	0.277	0.274	0.253
fim3_11	p	-0.087	0.077	0.171	0.171	0.146	0.105	0.088	0.048	0.097	-0.084	-0.137	-0.169	-0.145	0.006	-0.122
	p-value	0.444	0.502	0.132	0.132	0.200	0.359	0.439	0.675	0.397	0.465	0.233	0.139	0.207	0.957	0.285
fim3_12	p	-0.100	0.204	-0.066	-0.072	-0.066	-0.172	-0.049	0.067	-0.028	-0.024	0.077	-0.139	0.102	0.083	0.147
	p-value	0.379	0.071	0.563	0.526	0.564	0.130	0.669	0.557	0.810	0.836	0.503	0.227	0.375	0.469	0.197
fim3_13	p	-0.045	0.149	-0.070	0.016	-0.015	-0.165	-0.191	0.149	0.009	-0.002	0.024	0.016	-0.172	0.195	0.069
	p-value	0.697	0.190	0.540	0.888	0.894	0.147	0.092	0.191	0.941	0.985	0.832	0.890	0.132	0.085	0.543
fim3_14	p	-0.170	-0.020	0.213	0.174	0.039	0.148	0.228	0.046	-0.028	-0.002	-0.214	-0.109	-0.062	-0.145	-0.084
	p-value	0.133	0.859	0.060	0.126	0.734	0.193	0.044	0.689	0.806	0.985	0.060	0.342	0.591	0.201	0.463
fim3_15	p	-0.212	0.040	0.114	0.166	-0.023	0.114	0.118	0.029	-0.024	0.028	-0.356	-0.158	-0.125	-0.128	-0.109
	p-value	0.061	0.728	0.316	0.145	0.841	0.317	0.300	0.799	0.834	0.809	0.001	0.166	0.274	0.261	0.337

Tabela 63 Macierz korelacji moderatorów psychologicznych dla nagrania trzeciego.

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swiadomosc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielo znacznosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
fim4_01	p	-0.014	-0.026	0.136	0.212	0.173	0.078	-0.024	-0.116	-0.097	0.072	0.006	0.174	0.024	-0.072	-0.026
	p-value	0.901	0.822	0.230	0.059	0.125	0.490	0.832	0.306	0.395	0.529	0.958	0.125	0.834	0.528	0.820
fim4_02	p	0.135	-0.102	0.179	0.292	0.226	0.114	0.043	-0.135	-0.038	0.028	0.027	0.259	0.135	-0.067	-0.064
	p-value	0.234	0.369	0.112	0.009	0.044	0.315	0.705	0.231	0.742	0.805	0.810	0.021	0.234	0.554	0.571
fim4_03	p	-0.010	-0.011	0.154	0.326	0.153	0.080	0.017	-0.089	-0.124	0.007	-0.045	0.140	0.004	-0.116	-0.152
	p-value	0.932	0.920	0.174	0.003	0.176	0.479	0.883	0.431	0.275	0.949	0.692	0.218	0.973	0.306	0.177
fim4_04	p	0.036	-0.024	0.114	0.287	0.116	-0.000	-0.072	-0.169	-0.014	-0.054	-0.036	0.108	0.044	-0.094	-0.096
	p-value	0.754	0.834	0.313	0.010	0.304	1.000	0.524	0.135	0.902	0.636	0.750	0.344	0.703	0.407	0.396
fim4_05	p	-0.111	-0.044	0.182	0.265	0.088	0.070	-0.040	-0.249	0.115	-0.013	0.025	0.111	0.075	-0.171	-0.143
	p-value	0.328	0.695	0.107	0.018	0.439	0.536	0.725	0.026	0.314	0.912	0.829	0.329	0.509	0.130	0.205
fim4_06	p	-0.006	-0.022	0.136	0.331	0.111	0.025	-0.097	-0.097	-0.078	-0.067	-0.068	0.118	0.059	-0.088	-0.008
	p-value	0.961	0.847	0.228	0.003	0.329	0.825	0.391	0.390	0.492	0.599	0.553	0.299	0.604	0.437	0.943
fim4_07	p	0.037	0.058	-0.206	-0.201	0.004	0.031	0.046	-0.057	-0.024	0.086	-0.233	-0.193	0.054	-0.091	0.015
	p-value	0.744	0.611	0.066	0.074	0.975	0.784	0.688	0.615	0.834	0.450	0.039	0.088	0.634	0.420	0.896
fim4_08	p	0.029	0.278	-0.007	0.104	-0.018	-0.310	-0.124	0.080	-0.396	0.032	0.029	0.089	-0.017	0.059	0.120
	p-value	0.798	0.013	0.950	0.359	0.874	0.005	0.274	0.479	< .001	0.780	0.803	0.434	0.879	0.603	0.289
fim4_09	p	0.061	0.251	0.074	0.185	0.078	-0.170	-0.110	0.129	-0.364	0.099	0.198	0.200	0.048	0.072	0.176
	p-value	0.591	0.025	0.512	0.101	0.489	0.132	0.330	0.254	< .001	0.386	0.080	0.077	0.673	0.526	0.119
fim4_10	p	0.044	0.105	0.202	0.308	0.204	-0.011	-0.058	-0.146	-0.083	0.038	0.042	0.185	0.051	-0.074	-0.086
	p-value	0.699	0.355	0.073	0.005	0.070	0.923	0.610	0.196	0.469	0.741	0.713	0.103	0.658	0.512	0.449
fim4_11	p	-0.156	0.152	0.231	0.245	0.159	0.001	-0.135	-0.029	0.250	-0.023	0.116	0.044	0.068	-0.040	0.021
	p-value	0.167	0.178	0.039	0.028	0.159	0.991	0.231	0.800	0.026	0.844	0.310	0.703	0.550	0.725	0.855
fim4_12	p	-0.103	0.238	-0.006	0.185	0.131	-0.040	-0.036	0.021	-0.146	0.302	0.089	0.160	-0.002	0.021	0.120
	p-value	0.363	0.033	0.961	0.101	0.247	0.722	0.754	0.850	0.200	0.007	0.436	0.160	0.988	0.855	0.288
fim4_13	p	-0.117	0.115	0.162	0.204	0.097	-0.134	-0.280	0.004	0.205	-0.081	0.204	0.078	0.068	0.008	0.076
	p-value	0.300	0.311	0.151	0.069	0.393	0.238	0.012	0.973	0.071	0.480	0.071	0.497	0.552	0.941	0.502
fim4_14	p	0.099	0.033	0.334	0.267	0.255	0.108	-0.026	-0.155	0.081	-0.030	-0.009	0.094	0.046	-0.179	-0.185
	p-value	0.385	0.769	0.002	0.017	0.023	0.342	0.816	0.169	0.480	0.793	0.935	0.411	0.687	0.112	0.101
fim4_15	p	0.074	-0.050	0.218	0.258	0.120	0.183	0.063	-0.169	0.031	0.039	-0.136	0.112	0.086	-0.234	-0.219
	p-value	0.512	0.659	0.052	0.021	0.289	0.104	0.581	0.133	0.786	0.731	0.231	0.324	0.453	0.037	0.051

Tabela 64 Macierz korelacji moderatorów psychologicznych dla nagrania czwartego.

Tabela 65 Macierz korelacji moderatorów psychologicznych dla nagrania piątego.

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swietosc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielo znacznosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
film5_01	p	-0.113	0.260	0.065	0.227	0.108	-0.053	-0.017	-0.003	-0.138	0.005	-0.010	0.093	-0.064	0.010	0.061
	p-value	0.318	0.020	0.568	0.043	0.339	0.642	0.880	0.980	0.224	0.962	0.930	0.417	0.574	0.933	0.591
film5_02	p	-0.143	0.161	0.034	0.211	0.075	-0.099	-0.140	-0.071	-0.089	0.014	-0.023	0.144	-0.084	-0.061	0.067
	p-value	0.207	0.153	0.765	0.061	0.511	0.384	0.216	0.530	0.437	0.900	0.838	0.207	0.460	0.594	0.553
film5_03	p	-0.174	0.076	0.111	0.319	0.078	-0.083	-0.118	-0.159	0.010	-0.046	-0.078	0.025	-0.104	-0.107	-0.041
	p-value	0.122	0.504	0.328	0.004	0.491	0.464	0.296	0.159	0.931	0.688	0.496	0.827	0.361	0.346	0.717
film5_04	p	-0.088	0.011	0.162	0.283	0.055	-0.022	-0.062	-0.066	0.040	-0.111	0.042	0.010	0.022	-0.007	-0.094
	p-value	0.439	0.923	0.151	0.011	0.628	0.849	0.585	0.558	0.724	0.329	0.713	0.932	0.849	0.950	0.408
film5_05	p	-0.053	-0.086	0.208	0.334	0.124	-0.014	-0.168	-0.151	0.160	-0.119	0.049	0.041	0.114	-0.086	-0.191
	p-value	0.643	0.450	0.065	0.002	0.273	0.899	0.136	0.182	0.159	0.296	0.669	0.723	0.317	0.448	0.090
film5_06	p	-0.072	-0.014	0.232	0.401	0.132	-0.046	-0.146	-0.209	0.027	-0.154	0.003	0.076	0.064	-0.127	-0.165
	p-value	0.528	0.903	0.038	< .001	0.243	0.685	0.196	0.063	0.810	0.175	0.976	0.505	0.577	0.262	0.144
film5_07	p	-0.088	0.186	-0.134	-0.012	0.039	0.127	0.061	-0.029	0.044	0.009	-0.064	-0.100	-0.147	0.003	0.100
	p-value	0.439	0.098	0.236	0.913	0.729	0.261	0.594	0.799	0.699	0.935	0.576	0.378	0.196	0.976	0.378
film5_08	p	-0.071	0.299	0.084	0.273	0.102	0.063	0.137	-0.113	-0.086	-0.009	0.004	0.053	-0.013	-0.015	0.094
	p-value	0.533	0.007	0.461	0.014	0.369	0.578	0.225	0.319	0.453	0.941	0.972	0.645	0.910	0.897	0.405
film5_09	p	0.055	0.221	0.025	0.110	-0.012	-0.127	-0.038	0.038	-0.189	0.113	0.043	0.102	-0.024	-0.023	0.023
	p-value	0.629	0.049	0.824	0.331	0.915	0.261	0.736	0.737	0.095	0.322	0.707	0.369	0.835	0.837	0.842
film5_10	p	-0.177	0.143	0.112	0.191	0.003	-0.208	-0.129	0.006	-0.051	-0.152	0.079	-0.031	0.028	0.103	0.015
	p-value	0.116	0.205	0.322	0.090	0.981	0.064	0.254	0.955	0.657	0.181	0.490	0.786	0.809	0.361	0.897
film5_11	p	-0.127	0.041	0.318	0.240	0.115	0.017	-0.075	0.046	0.119	-0.113	0.046	-0.033	0.148	-0.037	-0.111
	p-value	0.262	0.718	0.004	0.032	0.308	0.878	0.511	0.684	0.297	0.320	0.689	0.775	0.193	0.741	0.325
film5_12	p	-0.065	0.296	0.101	0.079	0.113	-0.024	0.019	-0.148	-0.075	0.179	0.026	0.050	-0.002	-0.040	0.068
	p-value	0.568	0.008	0.372	0.487	0.319	0.833	0.865	0.191	0.511	0.114	0.822	0.662	0.989	0.726	0.548
film5_13	p	-0.113	0.036	0.194	0.238	0.169	-0.053	-0.208	-0.048	0.227	-0.072	0.212	0.057	0.112	-0.040	0.002
	p-value	0.320	0.753	0.085	0.033	0.135	0.639	0.065	0.671	0.044	0.530	0.060	0.618	0.327	0.723	0.987
film5_14	p	0.006	-0.065	0.140	0.184	0.081	-0.005	-0.138	-0.144	0.053	-0.115	-0.067	-0.024	0.106	-0.093	-0.167
	p-value	0.959	0.564	0.216	0.101	0.473	0.965	0.224	0.204	0.643	0.314	0.557	0.831	0.351	0.411	0.138
film5_15	p	0.061	-0.100	0.166	0.222	0.087	0.091	-0.076	-0.142	0.086	-0.041	-0.123	-0.094	0.184	-0.177	-0.187
	p-value	0.590	0.376	0.142	0.048	0.441	0.422	0.500	0.208	0.452	0.721	0.281	0.410	0.104	0.117	0.097

		IMPULSYWNOSC_samokontrola	IMPULSYWNOSC_impulsywnosc_w_dzialaniu	MFQ_V_swiet osc_degradacja	MFQ_IV_autorytet_kwestionowanie_wladzy	MFQ_III_lojalnosc_zdrada	MFQ_II_sprawiedliwosc_oszustwo	MFQ_I_troska_krzywda	SES	PDP_V_zdecydowanie	PDP_IV_zamknietosc_umyslowa	PDP_III_nietolerancja_wielo znacznosci	PDP_II_preferowanie_przewidywalnosci	PDP_I_preferowanie_porzadku	DEPRESJA	LEK
film6_01	p	0.037	0.166	0.052	0.086	-0.007	-0.296	-0.159	-0.020	-0.049	-0.021	0.134	0.037	0.112	0.053	0.040
	p-value	0.747	0.143	0.650	0.449	0.951	0.008	0.162	0.862	0.671	0.855	0.242	0.746	0.331	0.642	0.728
film6_02	p	-0.024	0.112	0.067	0.085	-0.050	-0.336	-0.189	-0.010	-0.008	-0.066	0.159	-0.048	0.008	0.066	0.032
	p-value	0.834	0.325	0.558	0.455	0.664	0.002	0.095	0.930	0.942	0.568	0.166	0.679	0.948	0.564	0.779
film6_03	p	0.045	0.040	0.051	0.196	0.030	-0.294	-0.219	-0.078	0.108	-0.129	0.102	-0.003	0.039	-0.016	0.017
	p-value	0.697	0.728	0.656	0.083	0.793	0.009	0.053	0.495	0.347	0.259	0.374	0.978	0.732	0.892	0.883
film6_04	p	-0.004	0.065	0.050	0.189	0.020	-0.336	-0.353	-0.095	0.103	-0.165	0.108	0.063	0.032	-0.015	0.050
	p-value	0.971	0.568	0.663	0.096	0.864	0.002	0.001	0.403	0.370	0.150	0.345	0.582	0.780	0.894	0.661
film6_05	p	-0.080	0.136	0.102	0.221	-0.035	-0.309	-0.290	0.012	0.042	-0.130	0.137	0.006	-0.023	0.082	0.083
	p-value	0.481	0.234	0.370	0.051	0.761	0.006	0.010	0.916	0.715	0.257	0.233	0.958	0.843	0.472	0.466
film6_06	p	-0.151	0.100	0.010	0.206	-0.032	-0.307	-0.212	-0.001	0.068	-0.129	0.135	0.018	0.026	0.060	0.101
	p-value	0.185	0.380	0.928	0.069	0.780	0.006	0.061	0.990	0.552	0.261	0.238	0.876	0.823	0.600	0.377
film6_07	p	-0.009	0.026	-0.134	-0.170	0.035	0.091	-0.018	0.082	-0.053	0.149	-0.035	-0.016	0.028	0.112	0.094
	p-value	0.940	0.821	0.239	0.134	0.760	0.426	0.874	0.470	0.647	0.191	0.761	0.890	0.811	0.326	0.409
film6_08	p	-0.013	0.227	-0.016	0.070	0.031	-0.372	-0.201	0.058	-0.096	-0.116	0.202	-0.012	0.124	0.151	0.162
	p-value	0.908	0.044	0.890	0.539	0.788	< .001	0.076	0.612	0.405	0.310	0.076	0.916	0.280	0.185	0.155
film6_09	p	0.018	0.201	0.047	0.150	0.063	-0.302	-0.167	-0.035	-0.044	-0.077	0.225	0.069	0.155	0.123	0.129
	p-value	0.876	0.076	0.681	0.187	0.580	0.007	0.141	0.762	0.700	0.504	0.048	0.547	0.175	0.282	0.257
film6_10	p	-0.064	0.122	-0.025	0.142	-0.017	-0.316	-0.249	-0.100	0.077	-0.114	0.173	0.123	0.138	0.053	0.139
	p-value	0.575	0.286	0.829	0.211	0.883	0.005	0.027	0.382	0.501	0.322	0.131	0.284	0.228	0.646	0.221
film6_11	p	-0.039	0.191	0.009	0.046	-0.023	-0.373	-0.264	0.023	0.028	-0.112	0.082	-0.111	-0.015	0.055	0.126
	p-value	0.732	0.092	0.935	0.686	0.840	< .001	0.019	0.841	0.806	0.330	0.476	0.334	0.898	0.630	0.270
film6_12	p	0.071	0.131	0.048	-0.005	0.077	-0.314	-0.204	-0.149	-0.030	0.002	0.280	0.157	0.287	0.090	0.130
	p-value	0.535	0.250	0.674	0.964	0.499	0.005	0.072	0.189	0.797	0.989	0.013	0.170	0.011	0.432	0.252
film6_13	p	-0.105	0.308	-0.063	-0.029	-0.135	-0.437	-0.319	0.035	-0.037	-0.074	0.034	-0.140	-0.065	0.075	0.167
	p-value	0.358	0.006	0.579	0.799	0.236	< .001	0.004	0.759	0.746	0.521	0.770	0.222	0.569	0.510	0.141
film6_14	p	-0.188	0.235	0.056	0.206	-0.010	-0.331	-0.296	0.030	0.083	-0.152	0.210	-0.000	-0.080	0.108	0.162
	p-value	0.097	0.037	0.625	0.068	0.930	0.003	0.008	0.795	0.471	0.185	0.065	0.999	0.485	0.342	0.154
film6_15	p	-0.151	0.155	0.091	0.185	-0.013	-0.286	-0.303	-0.051	0.124	-0.129	0.178	0.064	-0.037	0.025	0.121
	p-value	0.184	0.173	0.423	0.102	0.913	0.011	0.007	0.656	0.280	0.261	0.118	0.578	0.746	0.824	0.288

Tabela 66 Macierz korelacji moderatorów psychologicznych dla nagrania szóstego.

6.4.1 Impulsywność

Impulsywność to cecha osobowości lub zachowania, która charakteryzuje się skłonnością do podejmowania działań bez wcześniejszego przemyślenia konsekwencji lub bez odpowiedniej kontroli nad impulsem. Osoby impulsywne często podejmują szybkie decyzje lub reagują na bodźce bez zastanowienia się nad długoterminowymi konsekwencjami. W niniejszym podpodrozdziale analizie poddana została korelacja impulsywności z odpowiedziami na zadane pod filmami pytania.

Jak wynika z powyższych tabel, w przypadku obu filmów deepfake (1 i 6) i pytania 8 (W jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycyjne innych osób oglądających go?) zaobserwowano pozytywną korelację z impulsywnością w działaniu. Podobną zależność zaobserwowano również z pozostałymi filmami, prezentującymi zarówno nieznane osoby (nagrania 2 i 3) jak i średnio znanych celebrytów (filmy 4 i 5). Pokazuje nam to, iż osoby, które są bardziej impulsywne w działaniu uważają, że inne osoby będą podejmowały decyzje o swoich finansach na podstawie treści marketingowych w postaci filmów niezależnie od tego czy te filmy wytworzone zostały przy pomocy technologii deepfake, czy nie, i niezależnie od tego czy pokazują znane czy nieznane osoby. Wnioskować można, iż najprawdopodobniej osoby te projektują swoje cechy jednostki (impulsywność).

Pozytywną korelację z impulsywnością w działaniu zaobserwowano również w przypadku obu filmów deepfake (1 i 6) dla pytania 13 (Czy znasz osobę wyświetlaną na nagraniu?). Świadczyć to może o tym, iż osoby z podwyższoną impulsywnością lepiej rozpoznają osoby bez posiadania obrazu pełnego spektrum ich cech indywidualnych (na nagraniach dźwięk głosu nie był podrabiany).

Oprócz tego wykryto pozytywną korelację z impulsywnością w działaniu dla pierwszego z filmów deepfake (film 1) w pytaniu 9 (W jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?) i 12 (Jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?). Wnioskować można, iż osoby bardziej impulsywne w działaniu skłonniejsze są do zainwestowania swoich środków po obejrzeniu tego nagrania. Ponadto, osoby bardziej impulsywne w działaniu silniej oceniają negatywny wpływ filmu deepfake na odbiór prezentowanego twórcy przez znajomych / obserwujących. Jest to zgodne z pierwotnymi założeniami, przyjętymi w korelacji z moderatorem impulsywności. Spodziewano się, że osoby z podwyższonym wskaźnikiem społecznym

Impulsywność, będą bardziej podatne na manipulację i podejmowanie decyzji finansowej bez dłuższego zastanowienia.

6.4.2 Kody moralne

Zastosowany w pracy kwestionariusz „Moral Foundations Questionnaire" ocenia pięć głównych fundamentów moralnych:

Troska/Krzywda – ten kod moralny odnosi się do unikania krzywdzenia innych i troszczenia się o ich dobrostan. Koncentruje się na uczuciach troski, empatii i współczucia dla innych osób, szczególnie tych w potrzebie. Osoby o podwyższonym wyniku w tym wymiarze zwykle kładą większy nacisk na opiekę i sprawiedliwość społeczną. Są one wrażliwe na cierpienie innych i chętnie pomagają potrzebującym.

Sprawiedliwość/Oszustwo – skupia się na uczciwym traktowaniu wszystkich ludzi i dążeniu do równości. Osoby z podwyższonym wynikiem zazwyczaj przypisują dużą wagę równości i sprawiedliwości społecznej. Są one wyczulone na niesprawiedliwość i sprzeciwiają się nierównemu traktowaniu.

Lojalność/Zdrada – dotyczy lojalności wobec grupy i ochrony jej interesów. Osoby o wysokim wyniku w tej dziedzinie zazwyczaj cenią lojalność wobec swojej grupy społecznej czy narodowej. Są one skłonne do współpracy z członkami swojej grupy i bronienia jej przed zagrożeniami.

Autorytet/Uległość – ten kod moralny koncentruje się na szacunku dla władzy i autorytetu. Osoby z wyższym wynikiem zazwyczaj przywiązują większą wagę do poszanowania dla osób i instytucji o wyższym statusie społecznym. Są one skłonne do posłuszeństwa regułom i podporządkowywania się osobom na wyższych stanowiskach.

Świętość/Upodlenie – to wskaźnik czystości i świętości. Osoby o wysokim poziomie w tej dziedzinie zazwyczaj kładą nacisk na moralną czystość i unikanie zachowań uznawanych za grzeszne lub niemoralne. Są one wrażliwsze na brud i skażenie, zarówno fizyczne, jak i moralne.

W analizie krzyżowej przeanalizowano odpowiedzi na poszczególne pytania pod kątem wykrycia wskaźników charakterystycznych dla każdej z grupy nagrań. Patrząc na grupy, z pewnością zauważyć można intensywną negatywną korelację MFQ sprawiedliwość – oszustwo w obu filmach deepfake. Film pierwszy to pytania 1, 2, 3, 4, 6, 9, 10, 12, 13, 14 i 15. Dla drugiego z filmów deepfake (film 6) to pytania wszystkie

poza siódmym. Korelują one ujemnie prawie z każdą odpowiedzią na pytania dotyczące filmu deepfake. Podobne korelacje praktycznie w ogóle nie występują w przypadku pozostałych nagrań. W przypadku filmów deepfake należy więc stwierdzić, iż im wyższy jest wynik MFQ sprawiedliwość – oszustwo, tym niższe są wartości odpowiedzi na pytanie.

Wyniki te sugerują, że osoby o wyższej ocenie MFQ sprawiedliwość – oszustwo są bardziej wyczulone na manipulacje przedstawione w filmach deepfake. Osoby te mogą być bardziej skłonne do negatywnej oceny nagrania i osoby przedstawionej w filmie, a także do kwestionowania wiarygodności informacji w nim zawartych. Wnioskować można, iż osoby o wyższej ocenie MFQ sprawiedliwość – oszustwo są bardziej skłonne do negatywnej oceny treści filmów deepfake.

Ponadto ciekawą obserwacją może być pozytywna korelacja MFQ autorytet – uległość występująca w przypadku części odpowiedzi udzielonych dla filmów 1 (deepfake) oraz 4 i 5 (średnio znani influencerzy). Świadczy to o tym, że im bardziej ktoś szanuje autorytet, tym wyższe zaznaczał odpowiedzi, zwłaszcza w filmach znanych osób oraz w przypadku deepfake z aktorką. Wyniki te sugerują, że osoby o wyższej ocenie MFQ autorytet – uległość są bardziej podatne na wpływ autorytetów. Osoby te mogą być bardziej skłonne do zaufania informacjom przekazywanym przez osoby, które postrzegają jako autorytety, a także do naśladowania ich zachowań i opinii.

W filmach prezentujących średnio znanych influencerów (nr 4 i 5), jak nagraniu deepfake (nr 1), w przypadku odpowiedzi na kilka pytań (pytania 5, 6, 7, 14), zaobserwować można również pozytywną korelację wartości MFQ świętości – upodlenia. Sugerować to może, że im ktoś przykłada większą wagę do własnego wyglądu, tym bardziej radykalnie ocenia wiarygodność influencerów pod kątem ich prezencji.

Pozostałe obserwacje wykazywały albo bardzo słabą korelację (wskazującą na brak związku) albo nie wykazywały jej w ogóle.

6.4.3 Potrzeba Poznawczego Domknięcia

Zastosowany w pracy kwestionariusz Potrzeby Poznawczego Domknięcia dotyczy pięciu różnych aspektów PDP. Są to:

Preferowanie porządku – dotyczy skłonności do organizowania otoczenia w uporządkowany sposób i preferowanie zorganizowanych środowisk.

Preferowanie przewidywalności – dotyczy pragnienia stabilnych, niezmiennych warunków i unikania niespodzianek, niepewności. Jest to również wzmożone dążenie do przewidywania przyszłych wydarzeń.

Nietolerancja wieloznaczności – dyskomfort związany z niejednoznacznymi lub niepewnymi sytuacjami i informacjami. Często objawia się jako odczuwanie dyskomfortu w obliczu niejasnych lub niekompletnych informacji.

Zamkniętość umysłowa – to skłonność do ignorowania informacji, które są sprzeczne z już posiadanymi przekonaniami. Oporność na zmianę poglądów i akceptację nowych informacji.

Zdecydowanie – szybkie podejmowanie decyzji i niechęć do ich ponownego rozważania. Również objawia się jako skłonność do szybkiego formułowania opinii i trwania przy nich.

Wyłącznie dla nagrań deepfake i pytania dwunastego (Jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?) zaobserwowano niską, ale pozytywną korelację względem moderatora Potrzeby Poznawczego Domknięcia – nietolerancja wieloznaczności. Brak tej korelacji w przypadku pozostałych materiałów sugeruje, że specyficzne cechy filmów deepfake, takie jak ich niejednoznaczność i potencjalne wprowadzenie w błąd, wywołują u osób o wysokiej nietolerancji wieloznaczności silniejsze reakcje. Mogą one świadczyć o tym, iż osoby te są bardziej skłonne do wyciągania szybkich wniosków i formułowania opinii na podstawie niepełnych lub niejednoznacznych informacji. Ponadto filmy deepfake, ze względu na swoją sztuczną naturę, mogą być bardziej podatne na nadinterpretację przez osoby o silnej PDP nietolerancja wieloznaczności, co prowadzi do wyolbrzymionego postrzegania ich wpływu na odbiór aktorów przez innych.

Podobną zależność zaobserwowano z odpowiedziami na pytanie dziewiąte (w jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?). Oznacza to, iż PDP nietolerancja wieloznaczności jest pozytywnie skorelowana ze skłonnością do inwestowania po obejrzeniu filmów typu deepfake, lecz nie po obejrzeniu filmów autentycznych. Potwierdza to wnioski wysnute we wcześniejszym akapicie i może sugerować, iż osoby o wysokiej nietolerancji wieloznaczności są bardziej skłonne postrzegać filmy deepfake jako bardziej manipulacyjne oraz uznawać ich wpływ na decyzje inwestycyjne innych. To może

wynikać z ich potrzeby jednoznacznych i pewnych informacji, które są zaburzone przez niejednoznaczność filmów deepfake i ich oszukańczy charakter.

6.4.4 Pozostałe moderatory

Pozostałe zastosowane w pracy moderatory to moderatory samooceny, depresji i lęku. Pomimo starań, nie zaobserwowano więcej zależności. Pojedyncze korelacje są zbyt słabe lub zbyt jednostkowe (przypadkowe) by mówić o efekcie i poddawać je dalszej analizie.

Wstępnie spodziewano się wyników świadczących o tym, że im osoba jest bardziej lękowa tym będzie wskazywać niższe wyniki, bojąc się inwestować na nieznanej platformie. W praktyce nie zaobserwowano jednak żadnych istotnych korelacji z depresyjnością czy lękiem.

Nie zaobserwowano również żadnych istotnych statystycznie korelacji z poczuciem własnej wartości (dla żadnego z filmów). Może to sugerować, iż decyzje finansowe podejmowane na podstawie reklam marketingowych są niezależne od poczucia wartości danej osoby. Wstępnie spodziewano się, iż osoby mające mniejsze poczucie własnej wartości będą chciały ją wzmocnić poprzez chętniejsze inwestowanie, dla tezy tej nie znaleziono potwierdzenia.

6.5 Praktyczny wymiar badań dla bezpieczeństwa narodowego

Niniejszy podrozdział powstał w odpowiedzi na rosnące zapotrzebowanie na strategie i polityki przeciwdziałania ryzykom związanym z rozwojem AI, a zwłaszcza technologii deepfake. Dotychczas teoretyzowane zagrożenia, z biegiem czasu stają się coraz bardziej realne. Jak zauważono w niniejszej publikacji, zagrożenie ze strony technologii deepfake stanowi obecnie jeden z głównych tematów prac naukowych, seminariów i badań w wielu dziedzinach. Poruszane jest to wśród dziedzin takich jak nauki o bezpieczeństwie, prawnych, informatyka techniczna i telekomunikacja czy psychologia. W poprzednich rozdziałach prowadzone były rozważania dotyczące wyników badania laboratoryjnego i analizy danych zastanych. Narzędzia i techniki przeciwdziałania dezinformacji zaproponowane w niniejszym podrozdziale stanowią wynik analiz materiałów empirycznych opisanych w poprzednich rozdziałach. Uwzględniona została konieczność utworzenia gotowej strategii działania, ze wskazaniem elementów krytycznych dla bezpieczeństwa narodowego.

Przedmiotem rozważań niniejszego podrozdziału jest model przeciwdziałania nieprawdziwym mediom audiowizualnym oraz minimalizacja podatności społeczeństwa na rozwój dezinformacji prowadzonej przy pomocy technologii deepfake. W kontekście tego zagadnienia, pojęcie deepfake odnosi się do zaawansowanej technologii manipulacji treściami wideo, audio lub obu treści połączonych w jedno nagranie, które wyglądają i brzmią autentycznie, ale są w rzeczywistości spreparowane lub sfabrykowane. Celem analizy jest zaproponowanie rozwiązań mających na celu zwiększenie świadomości społecznej oraz uodpornienie jej na prawdopodobny nagły atak nieprawdziwych lub wprowadzających w błąd nagrań. Pytanie badawcze, które postawione sformowane zostało w pierwszym rozdziale, dla niniejszego podrozdziału brzmi „jakie należy podjąć działania w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake?”. Zaproponowana na hipoteza badawcza, wynikająca z wstępnego rozpoznania tematu brzmi, że „aby ochronić się przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, należy podjąć działania zapobiegające rozprzestrzenianiu się takich materiałów, edukować społeczeństwo w zakresie rozpoznawania dezinformacji audiowizualnej oraz zapewnić odpowiednie narzędzia do weryfikacji prawdziwości takich materiałów”.

Dla usystematyzowania zarówno wiedzy jak i stosowanej terminologii oraz uporządkowania pracy w niniejszym podrozdziale, wykorzystano DISARM Framework

w wersji 1.4[269]. Są to opracowane przez grupę badaczy w 2019 roku, open-source'owe ramy opisujące modele dezinformacji. Ramy zostały opracowane w oparciu o najlepsze praktyki w zakresie globalnego cyberbezpieczeństwa i zachęcają do ujednolicania terminologii, udostępniania danych i analiz oraz koordynowanie wspólnych działań zwalczających dezinformację. Ramy zostały skonstruowane na podstawie zarówno historycznych, jak i hipotetycznych działań i technik stosowanych przez manipulatorów oraz reakcji obrońców. Celem było dostarczenie możliwie jak najbardziej kompleksowego zestawu znanych i przewidywanych zachowań manipulatorów i działań obrońców. W niniejszym podrozdziale opisane zostały potencjalne ataki dezinformacji deepfake, poszczególne etapy ich prowadzenia oraz proponowane strategie przeciwdziałania im, opracowane na podstawie niniejszego warsztatu.

Technologia deepfake w wybranym frameworku, zaznaczona została w trzech czynnościach, odbywających się w początkowej fazie ataku, nazwaną fazą przygotowywania (*prepare)*, a dokładnie w grupie taktyk tworzenia treści (*TA06 Develop Content*). Zgodnie z frameworkiem oraz niniejszą pracą, wyznaczone zostały 3 techniki: generowanie obrazów przy pomocy AI (*Develop AI-Generated Images – Deepfakes*), generowanie filmów przy pomocy AI (*Develop AI-Generated Videos – Deepfakes*) oraz generowanie dźwięku przy pomocy AI (*Develop AI-Generated Audio – Deepfakes*). Dla usystematyzowania działań, podzielono możliwe ataki na cztery kategorie: dezinformacyjne, oszukańcze, zastraszające lub kompromitujące daną osobę lub organizację. Każda z nich stanowi jedynie jeden z wielu elementów kampanii, której przebieg został przeanalizowany poniżej oraz zaproponowane zostały środki zaradcze celem zatrzymania poszczególnych jej etapów.

W przypadku działań dezinformacyjnych prowadzonych z wykorzystaniem technologii deepfake wyróżnić można szereg taktyk, w których technologia ta może zostać wykorzystana. Przede wszystkim nagrania lub obrazy mogą być przekazane i udostępnione przez nieświadome lub działające z premedytacją osoby publiczne (*Bait legitimate influencers*). Zgodnie z wynikami przeprowadzonego badania, udostępniane przez nie materiały cechują się powszechniejszą akceptacją względem nieznanych nagrań. Ponadto duża liczba obserwujących i algorytmy mediów społecznościowych, promujących wytwory influencerów, przekładają się na dużą oglądalność w krótkim czasie.

[269] Strona główna projektu DISARM Framework https://www.disarm.foundation/framework [dostęp: 30.03.2024].

Bliźniaczą taktyką jest tworzenie i korzystanie z kont fałszywych ekspertów (*Use fake experts*). Podczas tego procesu powstać mogą zarówno całkowicie nowe osoby jak i mogą to być próby podszycia się pod istniejące autorytety. Technologia deepfake może ułatwić kradzież ich tożsamości poprzez tworzenie fikcyjnych wypowiedzi lub pomóc zanonimizować własne nagrania (ukryć tożsamość osoby atakującej) w przypadku tworzenia zmyślonego auorytetu.

Inną taktyką dezinformacji jest ukrywanie materiału deepfake w częściowo prawdziwej treści (*Seed Kernel of truth*). Może być to zarówno fałszywa grafika umieszczona w początkowo prawdziwym artykule, jak i nagranie deepfake, którego fragmenty okazały się prawdziwe. Działania te przyczynić mogą się do podważenia narracji powszechnie przywoływanej w docelowej grupie odbiorców lub promować narrację mniej powszechną wśród docelowych odbiorców, ale preferowaną przez atakującego.

Dezinformacja deepfake zamieszczana może być w wielu miejscach. Materiały te mogą być prezentowane w formie reklamy (*Deliver Ads*), na profilach w mediach społecznościowych (*Social media*), w mediach tradycyjnych (*Traditional Media*), jako memy (*Share Memes*), treść posta (*Post Content*) czy też komentarz lub odpowiedź na dany temat (*Comment or Reply on Content*). Oprócz tego materiały dezinformacyjne deepfake można rozsyłać jako bezpośrednie wiadomości grupowych lub postów (*One-Way Direct Posting*) bez możliwości odpowiedzi na zamieszczane treści.

Zgodnie z frameworkiem, cele kampanii dezinformacyjnej prowadzonej z użyciem technologii deepfake, będą identyczne jak dla każdej innej operacji wprowadzania w błąd. Celem mogą być prowadzenie propagandy państwowej (*Facilitate State Propaganda*), zniekształcanie (*Distort*) lub rozpraszanie przekazu (*Distract*), a także dzielenie (*Divide*) społeczeństwa, wywołując skrajne reakcje na kontrowersyjen treści.

Wyróżnić należy również działania mające na celu zastraszenie jednostki lub danej grupy społecznej. Wówczas celem kampanii zastraszającej jest wywołanie przerażenia (*Dismay*) w ofierze ataku, tym sposobem nakłonienie jej do konkretnego działania lub zaniechania dotychczasowych czynności. Może być to zarówno grożenie krytykowi lub dziennikarzowi relacjonującym dane wydarzenia, jak również wywieranie wpływu na osoby publiczne. Dalszym celem zastraszania może być wywołanie reakcji u ofiary i nakłonienie jej do konkretnego działania lub zaniechania dotychczasowych czynności.

Technologia deepfake wykorzystywana może być do podszywania się pod osoby (*Create personas*) związane z ofiarą i informowanie ofiary o ich rzekomym wypadku lub

porwaniu przy pomocy nagrania przygotowanego wcześniej i przesłanego przez komunikator (*Use Encrypted Chat Apps*) lub rozmowy „na żywo" (*Audio Livestream*) nawet z włączoną kamerą i obrazem (*Video Livestream*).

Oprócz tego wyróżnić należy doxing (*Dox*) oraz grożenie nim (*Threaten to Dox*). W wypadku grożenia, technologia deepfake może służyć do wytworzenia fałszywych materiałów, które mogą zostać wykorzystane do szantażu jednostki lub danej grupy. Przy doxingu, treści deepfake mogą być wymieszane z prawdziwymi materiałami (*Seed Kernel of truth)*. Może to wpłynąć na trudność rozróżnienia prawdy od fałszu, a tym samym spotęgować nękania ofiary.

Zbliżonym do kampanii zastraszającej jest schemat działania prowadzący do kompromitacji danej osoby lub organizacji. Celem może być obniżenie reputacji przeciwnika (*Degrade Adversary*) lub zdyskredytowanie danych źródeł (*Discredit Credible Sources*).

Fałszywe materiały deepfake prezentować mogą kompromitujące daną jednostkę lub organizację materiały. Publikowane mogą być one zarówno na profilach podszywających się pod prawdziwe atakowanego celu (*Prepare Assets Impersonating Legitimate Entities*), jak i na kontach przejętych w wyniku cyberataku (*Compromise legitimate accounts*). Rozpowszechnianie treści nierzadko zachodzi poprzez udostępnianie przez anonimowe konta *(Create Anonymous Accounts)* oraz rozprowadzanie przez konta – trolle (*Trolls amplify and manipulate)*.

Ponadto materiały deepfake posłużyć mogą do podszywania się nie tylko pod osoby publiczne, ale również różne organizacje (*Prepare Assets Impersonating Legitimate Entities*). Prowadzić to może do tworzenia fałszywych zbiórek pieniędzy (*Conduct fundraising)* lub kampanii crowdfundingowych (*Conduct Crowdfunding Campaigns*), również na nieistniejące cele.

Technologia deepfake wykorzystywana może być również do zebrania materiału o ofierze. Posłużyć może na przykład do stworzenia profilu fikcyjnej postaci na portalu randkowym (*Dating Apps*), nawiązaniu relacji z ofiarą i wyciągnięcia od niej kompromitujących ją materiałów, służących następnie do szantażu bądź wymuszania okupu. Badane są również sytuacje w których przy pomocy filmu deepfake oszust tworzy konto bankowe na obcą osobę lub znając jej dane osobowe, przejmuje dostęp do jej konta.

Na wspomnienie zasługują ponadto jeszcze inne działania oszustów prowadzone z użyciem technologii deepfake. Zgodnie z wybranym frameworkiem, fałszywe materiały

udostępniane mogą być przez nich zarówno w formie złośliwego postu (*Post Content*), reklamy (*Deliver Ads*), czy jako komentarz lub odpowiedź do danej konwersacji (*Comment or Reply on Content*). Jak zaznaczono w niniejszej pracy, obecnie najczęściej wykorzystywanym kanałem do propagowania fałszywych nagrań deepfake są reklamy. Dzięki nim oszuści w łatwy sposób i niewielkim kosztem są w stanie dotrzeć do wybranej przez siebie grupy społecznej (*Purchase Targeted Advertisements*, oszustwo fałszywych inwestycji). Fałszywe posty mogą mieć charakter „clickbaitu" (*Create Clickbait*) i prowadzić do fałszywych stron wyłudzających poświadczenia do danego medium społecznościowego, a nawet banku.

Dla każdego z tych działań przygotować należy odpowiadające im reakcje. W zależności od przeprowadzanej kampanii oraz jej zakresu dobierane środki będą się od siebie różnić. Przede wszystkim znaczenie ma tu skala oraz istotność potencjalnych szkód. Poniżej przygotowano macierz istotności skutków wywołanych daną kampanią.

Macież istotności	Niski zasięg (1)	Średni zasięg (2)	Wysoki zasięg (3)
Niska ważność (1)	(1,1) Mała istotność	(1,2) Umiarkowana istotność	(1,3) Umiarkowana istotność
Średnia ważność (2)	(2,1) Umiarkowana istotność	(2,2) Znacząca istotność	(2,3) Wysoka istotność
Wysoka ważność (3)	(3,1) Znacząca istotność	(3,2) Wysoka istotność	(3,3) Bardzo wysoka istotność

Tabela 67Macierz istotności skutków. Opracowanie własne.

W zaprezentowanej macierzy istotności skutków, posłużono się dwuwymiarowym modelem analizy, który uwzględnia dwa kluczowe aspekty: istotność dotykanej grupy osób (skala ważności dotkniętej grupy) oraz liczbę osób dotkniętych atakiem (skala zasięgu ataku). Założono bowiem, iż większe znaczenie będzie miała kampania dotykająca grupę osób pełniących funkcje publiczne niż tę samą liczbę nieznanych osób. Również skala ma wpływ na istotność skutków. W przypadku ataku na część społeczeństwa, dotkliwość skutków będzie wielokrotnie większa niż tego samego ataku ukierunkowanego na jedną osobę.

Tworząc odpowiednią strategię działania, należy wpierw zdefiniować skalę istotności potencjalnych skutków i dopasować do każdej z nich odpowiednie techniki przeciwdziałania oraz ich intensywność. Pozwoli to uniknąć nagłej eskalacji zagrożenia, jak również uniemożliwi powstania sytuacji, gdzie przydzielone zasoby do jednego z ataków uniemożliwiają prawidłową reakcję na nowe zagrożenie.

Niska ważność to przede wszystkim grupy marginalizowane, które mają ograniczony wpływ na społeczeństwo lub politykę. Średnia ważność to osoby z pewnym poziomem wpływu, ale nie kluczowe dla struktur władzy czy opinii publicznej. Zaś wysoka ważność oznacza kluczowe dla istnienia państwowości grupy społeczne i zawodowe, takie jak niezależni dziennikarze, liderzy opinii, politycy czy pracownicy wysokiego szczebla instytucji rządowych.

Przyjęta skala zasięgu ataku określa jak wiele osób zostało dotkniętych kampanią dezinformacyjną. Niski zasięg definiuje niewielką liczbę osób, np. lokalną społeczność. Średni zasięg dotyczy większej grupy, np. mieszkańców miasta lub specyficznej grupy społecznej na poziomie regionalnym. Wysoki zasięg oznacza bardzo dużą grupę osób dotkniętą atakiem, np. mieszkańców całego kraju lub poszczególnych województw.

Celem wykazania istotności, w macierzy zastosowano pięciostopniową skalę. Pomimo tego, charakter działań umiarkowanej istotności (1,3) będzie różnił się w niewielkim stopniu od działań mających stanowić odpowiedź na kampanię o umiarkowanej istotności (2,1). Poniżej scharakteryzowano każde z pól matrycy.

(1,1) Mała istotność, to przede wszystkim kampania o małym zasięgu i dotycząca grupy o niskiej ważności. Tego typu kampanie mają ograniczony wpływ i mogą nie wymagać działań im przeciwdziałających. (1,2) Umiarkowana istotność to kampania o średnim zasięgu, ale dotykająca grupy o niskiej ważności. Z założenia wymaga pewnej uwagi, ale nie jest priorytetem. Również (1,3) umiarkowana istotność, czyli kampania o wysokim zasięgu, ale dotykająca grupy o niskiej ważności nie powinno stanowić priorytetu działań. Chociaż wiele osób jest zaangażowanych, ich wpływ na szersze społeczeństwo jest ograniczony. Podobnie dzieje się z (2,1) umiarkowana istotność, gdzie prowadzona kampania charakteryzuje się małym zasięgiem, ale dotyczy grupy o średniej ważności. Kampania ta może mieć lokalne lub specyficzne znaczenie i wymaga monitorowania.

Kampanie o znaczącej istotności to (2,2), czyli kampania o średnim zasięgu i dotycząca grupy o średniej ważności. Wymaga skoordynowanej reakcji, ponieważ może

mieć istotne skutki. Jest to również (3,1), gdzie kampania charakteryzuje się małym zasięgiem, ale dotyczy grupy o wysokiej ważności. Może mieć krytyczne znaczenie mimo ograniczonego zasięgu.

Wysoką istotność obserwujemy w dwóch polach – (2,3) charakteryzująca się wysokim zasięgiem, ale dotycząca grupy o średniej ważności. Drugim polem jest (3,2), gdzie prowadzona kampania jest o średnim zasięgu, ale dotyczy grupy o wysokiej ważności. Obserwacja i przeciwdziałanie skutkom jest niezwykle istotne dla zachowania struktury społecznej i politycznej.

Najwyższą wrażliwością cechuje się grupa (3,3) bardzo wysoka istotność. Jest to kampania o wysokim zasięgu, dotycząca grupy o wysokiej ważności. Atak skierowany na tę grupę stanowić powinien najwyższy priorytet działania. Skutki mogą być trudne do przewidzenia, a niekontrolowana sytuacja może być potencjalnie destabilizująca działanie państwa.

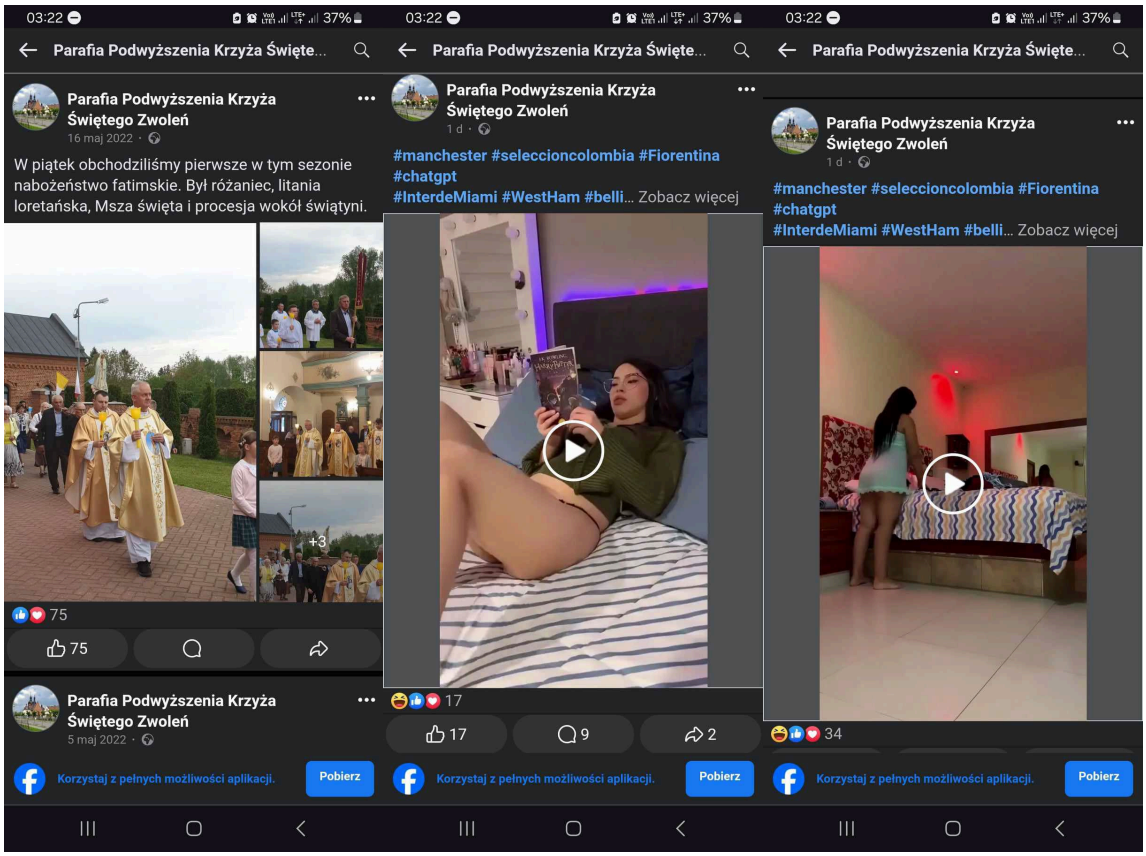
Przygotowując się na odparcie każdej z czterech zdefiniowanych kampanii wykorzystujących deepfake (kampanie dezinformacyjne, oszukańcze, zastraszające lub kompromitujące daną osobę lub organizację), podstawowy element stanowi plan strategii (*Plan Strategy*). W jego obszarze znajduje się przeciwdziałanie atakom oraz przygotowanie planu reagowania na dezinformację (*Have a disinformation response plan*).

Uniwersalnym, a przez to niezwykle istotnym elementem modelu przeciwdziałania atakom z użyciem technologii deepfake jest edukacja społeczna. Jest to element wzmacniający obronę przed każdą z czterech wymienionych powyżej kampanii wykorzystujących deepfake. Wykorzystany w niniejszej pracy framework identyfikuje zwłaszcza trzy główne elementy mające na celu podniesienie zdolności obronnych społeczeństwa. Pierwszym z nich jest szkolenie z zakresu inżynierii społecznej (*Counter social engineering training*), które obejmuje szkolenie w zakresie przeciwdziałania manipulacji, naukę weryfikowania faktów oraz edukację w zakresie zapobiegania phishingowi. Drugim z elementów jest tworzenie programów edukacyjnych dla influencerów (*find and train influencers*). W założeniu twórców frameworka, influencerom internetowym mającym duże zasięgi w mediach społecznościowych, oferowane powinno być wsparcie w formie szkoleń lub materiałów edukacyjnych. Wynika to z faktu, iż grupa ta posiada szerokie zasięgi w sieci, a często nie przykłada wagi do merytorycznej weryfikacji szerzonych informacji. Dotyczy to zwłaszcza kont budujących swą popularność na odtwarzaniu internetowych trendów i patologicznych, prowokujących zachowaniach.

Trzecią uniwersalną metodą obrony zdaje się być zwiększenie świadomości społecznej dotyczącej istnienia technologii deepfake oraz zagrożeń związanych z jej wykorzystaniem. Kampanie informacyjne, seminaria i programy edukacyjne mogą pomóc w zwiększeniu świadomości społeczeństwa na temat manipulacji treściami audiowizualnymi oraz pozytywnie wpłynąć na zdolności percepcji obywateli.

Elementem systemowego działania obronnego zwalczającego kampanie z wykorzystaniem technologii deepfake jest systematyczne wyszukiwanie i usuwanie podejrzanych kont (*remove suspicious accounts*) na platformach społecznościowych. Standardowe raportowanie fałszywych profili obejmuje wykrywanie przejętych kont i ponowne ich przydzielanie – jeśli to możliwe, z powrotem do pierwotnych właścicieli. Z punktu zagrożenia dezinformacją niezwykle istotne jest, aby platformy systematycznie monitorowały i usuwały oznaczone jak podejrzane konta (*Ensure that platforms are taking down flagged accounts)*. Jedną z potencjalnych metod zwiększania skali i skuteczności takich działań jest ustanawianie przez m.in. organizacje rządowe oficjalnych regulacji państwowych, chroniących w tej sferze interesy obywateli.

Rozwinięciem tej strategii ochrony przed kradzieżą tożsamości jest usuwanie starych, nieużywanych kont w mediach społecznościowych (*Delete old accounts / Remove unused social media accounts*), zarówno przez administrację platformy czy indywidualnych użytkowników. Istotnie może zmniejszyć to pulę kont dostępnych do przejęcia, zapobiegając eskalacji zjawiska podszywania się i rozpowszechniania fałszywych treści, również za pomocą modyfikacji deepfake.



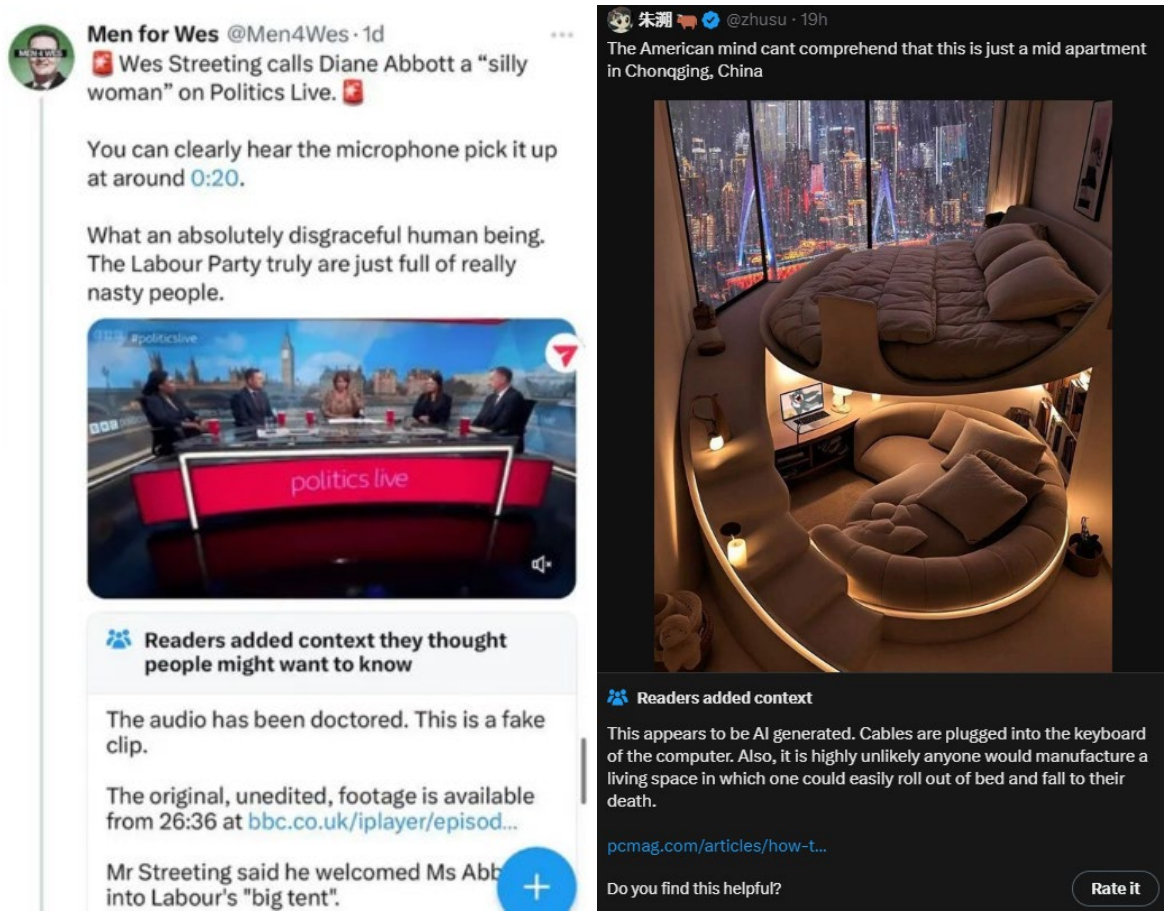
Grafika 25 Przejęty przez przestępców profil Parafii Podwyższenia Krzyża Świętego Zwoleń, na którym przez ponad dwa miesiące zamieszczano treści pornograficzne. Opracowanie własne.

Najpopularniejsze platformy społecznościowe (Facebook, X, Linkedin) udostępniają również możliwość weryfikacji profili (*third party verification for people*). W zależności od platformy rozwiązania te są w mniejszym bądź większym stopniu skomercjalizowane, jednak w swoim założeniu mają między innymi ograniczyć podszywanie się pod wpływowe osoby lub organizacje oraz przeciwdziałać powstawaniu fałszywych ekspertów.

Innym rozwiązaniem walki z dezinformacją, może być dodawanie przez platformę społecznościową etykiet ostrzegawczych, których umieszczenie odbywa się na podstawie głosów danej społeczności (*Platform adds warning label and decision point when sharing content*). Mechanizm ten z powodzeniem wdrożyła w połowie 2023 roku platforma społecznościowa X (wcześniej Twitter).

Do każdego postu, użytkownicy z rangą „*community notes*" dodawać mogą sprostowania nieprawdziwej treści. Po jej zamieszczeniu i akceptacji przez X, pozostali członkowie społeczności głosują nad publikacją sprostowania. Każdy z głosujących ocenia dany kontekst czytelniczy odpowiadając na następujące pytania: Czy notatka cytuje źródła wysokiej jakości? Czy jest łatwa do zrozumienia? Czy bezpośrednio odnosi się do tematu postu? Czy niesie ze sobą ważny przekaz? Czy jej język jest neutralny lub bezstronny?

Ponadto, w przypadku negatywnej oceny notatki, głosujący proszony jest o odpowiedź na następujące pytania: Co w tej notatce było nieprzydatnego? Czy źródła nie zostały uwzględnione lub są niewiarygodne? Czy źródła nie potwierdzają notatki? Czy notatka przekazuje nieprawidłowe informacje? Czy jest to opinia lub spekulacja? Czy są w niej literówki lub jej język jest niejasny? Czy pomija kluczowe elementy lub niesie ze sobą nieistotne punkty? Czy jej argumentacja jest stronnicza? Czy dana notatka nie jest potrzebna w tym poście? Czy dotyczy ona nękania lub stanowi nadużycie?



Grafika 26 Dwa przykłady dezinformacji prowadzonej na X z wykorzystaniem deepfake oraz AI. Oba zostały oznaczone kontekstem czytelników (Readers added context) prostujących nieprawdziwy przekaz. Opracowanie własne.

Inne podejście do weryfikacji fałszywych materiałów deepfake przyjęła firma Meta, wprowadzając na swoich głównych platformach społecznościowych – Facebook oraz Instagram – możliwość weryfikacji materiałów przez zweryfikowane ośrodki fact-checkingowe. Filmy czy też obrazy oznaczone przez nie jako fałszywe, dalej są prezentowane na platformach, jednak odpowiednio oznaczone jako „fałszywe informacje" wraz z dodatkowym komentarzem kontekstowym.

Grafika 27 Przykład oznaczenia fałszywego obrazu wytworzonego przy pomocy AI. Wpierw prezentowany materiał został zablurowany oraz oznaczony napisem „Fałszywe informacje", następnie po kliknięciu „zobacz post" wyświetlił się. Możliwe jest również uzyskanie informacji kontekstowej oraz nazwy organizacji zgłaszającej fałsz. Opracowanie własne.

Skutecznym sposobem przeciwdziałania dezinformacji tworzonej z wykorzystaniem deepfake jest identyfikacja i ujawnienie osób stojących za takimi działaniami oraz ich rzeczywistych intencji (*Expose actor and intentions*). W tym celu należy przeprowadzić dokładną analizę śladów cyfrowych, aby zidentyfikować osoby lub organizacje stojące za tworzeniem i rozpowszechnianiem fałszywych treści. Prezentacja niezbitych dowodów zaangażowania konkretnych podmiotów pozwala na zmniejszenie ich wiarygodności (*Provide proof of involvement*). Ważna jest przy tym transparentność źródeł – raporty wyjaśniające, kto jest odpowiedzialny za dezinformację, pomagają społeczeństwu zrozumieć motywy tych działań. Aby wzmocnić ujawnienie aktora, należy dostarczyć solidne dowody jego zaangażowania. Można to osiągnąć poprzez zebranie i przedstawienie dokumentacji dowodowej, takiej jak powiązania osobowe grupy, przechwycone wiadomości e-mail, lub inne materiały powiązane z tworzeniem i rozpowszechnianiem materiałów deepfake.

Zmniejszenie wiarygodności twórców dezinformacji można osiągnąć poprzez ujawnienie ich źródeł finansowania. Publiczne pokazanie, kto finansuje daną kampanię dezinformacyjną (*Denigrate the recipient/ project (of online funding*)), może znacząco osłabić jej oddziaływanie, zwłaszcza jeśli finansowanie pochodzi od grup ekstremistycznych lub państw nieprzyjaznych. Ważne jest również budowanie narracji, które podkreślają

nieetyczne postępowanie aktorów, takie jak manipulowanie opinią publiczną dla korzyści politycznych czy komercyjnych.

Marginalizacja i dyskredytacja grup ekstremistycznych publikujących dezinformację (*Marginalise and discredit extremist groups*) jest kolejnym etapem działania. Systematyczne monitorowanie działalności takich grup oraz raportowanie ich poczynań do organów ścigania i platform społecznościowych może prowadzić do szybszego usuwania wygenerowanych przez nie treści. Organizowanie kampanii zwalczających fałszywe informacje, które podważają wiarygodność i moralność grup ekstremistycznych, może również skutecznie ograniczyć ich wpływ.

Jako odpowiedź na pojedyncze akty kampanii wykorzystujących deepfake (zarówno kampanie dezinformacyjne, oszukańcze, zastraszające jak i kompromitujące daną osobę lub organizację), traktować można wydanie natychmiastowego komunikatu prostującego nieprawidłowe informacje przez osobę, której wizerunek bezprawnie wykorzystano (*Respected figure (influencer) disavows misinfo*).

W przypadku personalnego ataku np. W postaci karykaturalnych filmów deepfake rekomenduje się zamieszczenie humorystycznych reakcji (*Use humorous counter-narratives*) jako odpowiedzi na nie. Działania mające na celu usunięcie nagrań z Internetu lub próba ścigania prawnego osób stojących za karykaturą, przełożyć mogą się na tak zwany efekt streisand[270], co po pierwsze przełoży się na dużo szerszą propagację materiałów oraz uznane będzie przez atakujących jako sukces ich działania. Ponadto humor sam w sobie stanowić może narzędzie w demaskowaniu dezinformacji. Tworzenie memów i satyrycznych filmów, które ośmieszają i dekonstruują fałszywe narracje oraz aktorów, może skutecznie zmniejszyć ich wpływ. Angażowanie influencerów, komików i artystów w tworzenie treści, które w humorystyczny sposób podważają wiarygodność dezinformacji powstałej z wykorzystaniem deepfake, może przyczynić się do zwiększenia odporności społeczeństwa na tego rodzaju manipulacje.

W przypadku ataków o średnim albo wysokim poziomie ważności, jak również wysokim zasięgu, konieczne może być szybkie i szerokie ostrzeganie o zagrożeniu, np. Za pomocą systemów podobnych do Amber Alert (*Social media amber alert*) w mediach społecznościowych, bądź utworzonych do tego celu aplikacjach. W polskich warunkach

[270] Artykuł na blogu California Coast Line, opisujący działania w sprawie zdjęcia domu Barbary Streisand oraz kontrowersji z tym związanych https://www.californiacoastline.org/news/sjmerc5.html [dostęp:01.05.2024].

aplikację tę może stanowić mObywatel, przy wykorzystaniu, którego krótki komunikat otrzymywałaby duża część społeczeństwa korzystającego na co dzień z Internetu. Może być to również system SMS-owy, utworzony na wzór RCB alert. Wprowadzenie takiego systemu, który natychmiast informuje użytkowników mediów społecznościowych o próbie dezinformacji lub oszustwa, oraz współpraca z platformami społecznościowymi w celu automatycznego oznaczania i rozpowszechniania ostrzeżeń, może znacząco zwiększyć efektywność walki z dezinformacją.

6.6 Wnioski

Zgodnie z definicją przyjętą w niniejszej dysertacji, bezpieczeństwo strukturalne rozumie się jako ukierunkowanie działalności wszystkich instytucji życia społecznego, tak, aby ich działanie, a przede wszystkim jego efekty, gwarantowały bezpieczeństwo personalne. Zgodnie z tym w niniejszym rozdziale wpierw przeanalizowano wpływ pojedynczych nagrań na społeczeństwo oraz wybrane moderatory psychologiczne dotyczące jednostek, aby następnie zaproponować kompleksową strategię zwalczania dezinformacji i oszustw powstałych z wykorzystaniem technologii deepfake.

W niniejszym rozdziale jako jedynym sformułowano aż dwa pytania badawcze. Pierwsze z nich brzmi: „jaki wpływ mają zastosowane moderatory na odbiór dezinformacji deepfake, a przez to na bezpieczeństwo narodowe?". Odpowiadająca jej hipoteza jest następująca: osoby z podwyższonymi poszczególnymi wskaźnikami społecznymi są bardziej podatne na uleganie manipulacjom dezinformacji deepfake, co przyczynia się do zwiększenia ryzyka zagrożenia bezpieczeństwa narodowego. Weryfikacja tej hipotezy przeprowadzona została zwłaszcza w podrozdziale 6.4, ale również przy wsparciu wyników analiz materiałów empirycznych, opisanych w podrozdziałach 6.1, 6.2 oraz 6.3.

Drugie z pytań badawczych, na którego odpowiedź oparta jest na analizie przeprowadzonej nie tylko w piątym podrozdziale, ale przy pomocy materiału empirycznego całej dysertacji brzmi: „jakie należy podjąć działania w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake?". Odpowiedzią na to pytanie badawcze, ma być niniejsza hipoteza: aby ochronić się przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, należy podjąć działania zapobiegające rozprzestrzenianiu się takich materiałów, edukować społeczeństwo w zakresie rozpoznawania dezinformacji audiowizualnej oraz zapewnić odpowiednie narzędzia do weryfikacji prawdziwości takich materiałów.

Obie hipotezy zostały zweryfikowane. Uwzględniając wyniki badania przeprowadzonego w ramach niniejszej dysertacji, w celu ochrony przed dezinformacją realizowaną z wykorzystaniem technologii deepfake, na uwzględnienie zasługuje szereg obszarów zapewniających narzędzia zwalczania szkodliwych działań.

Jest to przede wszystkim dalszy rozwój technologii i postępujące wraz z tym jej możliwości. Niezbędna jest kontynuacja badań nad rozwijaniem skutecznych narzędzi i technologii służących wykrywaniu deepfake, przede wszystkim przez portale społecznościowe, mające największe możliwości przeciwdziałania dezinformacji i oszustwom. Zaawansowane algorytmy, sztuczna inteligencja i uczenie maszynowe mogą pomóc w identyfikacji fałszywych profili oraz udostępnianych przez nie treści wideo lub audio i umożliwiać szybką reakcję na dezinformację.

Do działań zapobiegawczych należy również bliska współpraca władz państwa z właścicielami platform społecznościowych, portali informacyjnych i innych podmiotów odpowiedzialnych za udostępnianie treści. Należy wywrzeć presję na te podmioty, celem stosowania przez nie skutecznych filtrów i algorytmów wykrywających deepfake. Bliska współpraca organów ścigania z omawianymi platformami przełożyć się również może na wykrywanie aktorów tworzących i udostępniających fałszywe treści. Organy państwowe powinny dbać o zachowanie ciągłego kanału komunikacji z branżą technologiczną. Przełożyć się to możne na zwiększoną presję względem międzynarodowych korporacji.

Konsekwentnie, przyjęcie odpowiednich regulacji, opracowanie strategii i polityki społecznej oraz ujednoliconych wytycznych sposobu reagowania na cyberataki na poziomie państwowym, może pomóc w ograniczeniu nielegalnego i szkodliwego użycia tej technologii w przypadku wykrycia zagrożenia atakiem dezinformacyjnym.

Dodatkowo, zapewnienie edukacji społeczeństwu w zakresie rozpoznawania nowych metod oszustw i prowadzenia dezinformacji, przełoży się na wzrost odporności społecznej. W odpowiedzi na rosnącą ilość nieprawdziwych lub wprowadzających odbiorcę w błąd nagrań oraz wciąż doskonalone metody manipulacji głosem i obrazem poprzez między innymi deepfakeprocess edukacji nie powinien być jednorazowy, a ciągły. Mogą się na niego składać kampanie informacyjne, seminaria i programy edukacyjne oraz materiały podnoszące kompetencje społeczne. Zwłaszcza do młodszych grup odbiorców trafić można z przekazem poprzez influencerów.

Ponadto wyuczenie zdolności weryfikacji źródeł informacji oraz krytycznego myślenia poskutkuje nie tylko wzmocnieniem odporności bezpieczeństwa strukturalnego,

lecz również bezpieczeństwa personalnego. Promowanie kultury krytycznego myślenia i weryfikacji źródeł informacji jest kluczowe w walce z dezinformacją. Społeczeństwo powinno być zachęcane do sprawdzania wiarygodności treści przed ich udostępnieniem lub opieraniem na nich swoich decyzji, na przykład inwestycyjnych.

Istotne jest również prowadzenie badań w obszarze współpracy międzynarodowej w zakresie zwalczania manipulacji informacyjnej. Tematem dyskusji na forum międzynarodowym powinna być wymiana doświadczeń związanych z nowymi rodzajami zagrożeń, ale również z najskuteczniejszymi praktykami ich zwalczania.

Kolejne obszary wymagające dalszych badań to skutki społeczno-gospodarcze internetowej dezinformacji i oszustw deepfake. Badanie konsekwencji rozpowszechniania fałszywych nagrań na zaufanie społeczne, procesy decyzyjne, rynki finansowe oraz stabilność polityczną może dostarczyć istotnych danych dotyczących charakteru i zasięgu tego zagrożenia

W dyskusji należy również uwzględnić podejście socjologiczne. Niezbędne jest dalsze prowadzenie badań socjologiczno-psychologicznych, celem pogłębiania wiedzy w zakresie wpływu cech osobowościowych jednostki na podatność na manipulację, w tym za pomocą treści audiowizualnych. Przełożyć się to może na nowe prace z obszaru bezpieczeństwa narodowego, gdzie odpowiednia analiza empiryczna dostarczonych danych korzystnie wpłynie na dalsze zgłębianie tematyki wpływu materiałów deepfake na rzeczywistość, a przede wszystkim pozwoli wyznaczyć dalsze związane z tym zagrożenia.

Zakończenie

W obliczu dynamicznego rozwoju technologicznego oraz powszechnego dostępu do narzędzi informatycznych kwestia autentyczności treści w mediach cyfrowych staje się coraz bardziej istotna. Technologia deepfake, będąca połączeniem zaawansowanej sztucznej inteligencji z algorytmami generatywnymi, wzbudza niepokój ze względu na swoją zdolność do tworzenia wiarygodnych manipulacji wideo i audio, mogących wprowadzać w błąd odbiorców oraz wpływać na ich decyzje. W kontekście narastających zagrożeń dla bezpieczeństwa narodowego kwestia potencjalnego wykorzystania technologii deepfake staje się niezwykle istotna dla społeczeństwa i instytucji państwowych. Ma to bezpośrednie przełożenie na cel główny niniejszej dysertacji, którym jest identyfikacja zagrożeń dla bezpieczeństwa narodowego, wynikających z rozwoju technologii deepfake. Cel praktyczny to wskazanie kierunków działania w celu ochrony przed dezinformacją oraz podnoszenie kompetencji obywatelskich w zakresie rozpoznawania i profilowanie psychologiczne osób podatnych na manipulacje.

Hipoteza badawcza postawiona w niniejszej pracy głosi, iż technologia deepfake pozwala na generowanie rzeczywistych nagrań, które mogą wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu. Badając możliwości i konsekwencje stosowania deepfake, autor pracy przeprowadził dogłębną analizę różnorodnych aspektów tego zjawiska, w tym jego potencjalne implikacje społeczne, polityczne, stwarzające zagrożenie dla bezpieczeństwa personalnego jak i strukturalnego. W trakcie prowadzonych badań przeprowadzono eksperyment laboratoryjny, analizę danych zastanych oraz badania terenowe (tworzenie nagrań deepfake), celem lepszego zrozumienia mechanizmów działania technologii oraz identyfikacji sposobów przeciwdziałania im.

Pytanie badawcze prezentowane w niniejszej pracy brzmi: czy technologia deepfake pozwala na generowanie rzeczywistych nagrań, mogących wpływać na decyzje osób je oglądających, a przez to zagrażać bezpieczeństwu narodowemu? Celem prawidłowej odpowiedzi na tak postawione pytanie, zadane zostało sześć pytań szczegółowych, mających za zadanie umożliwić głębsze zrozumienie problematyki i jej kompleksową analizę.

Podsumowując rozważania zawarte w niniejszej dysertacji, w pierwszej kolejności należy stwierdzić, iż pozwoliły one osiągnąć stawiane cele i zweryfikować pomocnicze hipotezy badawcze dotyczące przedmiotu badań.

Odnosząc się do postawionych w niniejszej pracy sześciu pomocniczych hipotez badawczych, należy stwierdzić, że podczas przeprowadzonych badań każda z nich została zweryfikowana. Jako punkt odniesienia ustalono status ontologiczny wiedzy na temat wpływu materiałów deepfake na bezpieczeństwo narodowe. Szczegółowo prześledzono proces ich powstawania, wpływ percepcji nagrań przez internautów na podejmowane przez nich decyzje oraz tworzone opinie uwzględniając wybrane moderatory psychologiczne. Ponadto stosując matrycę DISARM Framework przeanalizowano możliwe działania obronne w obszarze poszczególnych dziedzin.

Przede wszystkim wyniki analiz opisanych w pracy wskazują na złożoność i wielowymiarowość problemu, a także na konieczność dalszych badań interdyscyplinarnych, które pozwolą na pełniejsze zrozumienie omawianych zjawisk.

W rozdziale pierwszym przedstawione zostały założenia metodologiczne. Omówione zostały szczegółowo problemy i zmienne w procesie badawczym, hipotezy wymagające weryfikacji, cele pracy oraz sytuacja problemowa, czyli uzasadnienie podjęcia badań nad technologią deepfake i jej wpływem na bezpieczeństwo narodowe. Ponadto opisane zostały zastosowane w pracy teoretyczne i empiryczne metody badawcze. Do głównych empirycznych metod badań zaliczyć należy analizę danych zastanych oraz eksperyment badawczy, który stanowi główny element pracy.

Na podstawie analiz przeprowadzonych w części teoretycznej (rozdział drugi) określono istotę informacji jako jednego z najważniejszych czynników bezpieczeństwa narodowego w państwie demokratycznym. Ponadto przybliżono ustalenia psychologii społecznej oraz znane metody manipulacji w kontekście jej wpływu na poziom bezpieczeństwa narodowego. Usystematyzowano zebraną wiedzę, przybliżono siatkę pojęciową oraz zdefiniowano podłoże mechanizmów dezinformacyjnych. Na tej podstawie dokonano oceny statusu ontologicznego wiedzy na temat wpływu zmanipulowanych materiałów audiowizualnych na bezpieczeństwo narodowe oraz sformułowano elementy dalszych działań.

Wykazano, iż technologia deepfake niesie ze sobą nowe, potencjalne zagrożenia, a filmy deepfake mogą zostać wykorzystane do rozpowszechniania dezinformacji i propagandy, co może podważyć zaufanie do instytucji i zdestabilizować społeczeństwa.

Przeanalizowano również sprawy sądowe, gdzie nagrania deepfake używane były do tworzenia fałszywych dowodów.

Zwrócono również uwagę na trudności oceny rzeczywistego wpływu nagrań deepfake, ze względu na brak badań w tym zakresie lub ich ograniczony zasięg. Brak jest również jakichkolwiek statystyk związanych z ilością tworzonych filmów deepfake, jak i ich wykorzystaniem w celach oszustwa czy dezinformacji.

W rozdziale poruszono w związku z tym konieczność tworzenia nowych, interdyscyplinarnych badań z tego obszaru. Wykazano, iż do dalszych analiz, niezbędna jest współpraca specjalistów z takich dyscyplin jak nauki o bezpieczeństwie, informatyka, psychologia, politologia czy prawo.

Następnym elementem niniejszej pracy, mieszczącym się w rozdziale trzecim, było przygotowanie do dalszych badań fałszywych nagrań deepfake oraz szczegółowe opisanie tego procesu. Zgodnie z przyjętymi założeniami, wszystkie nagrania utworzone zostały przez autora pracy. Każdy z kolejno wykonywanych kroków został szczegółowo przeanalizowany i opisany, wraz z komentarzami autora. Zdobyte doświadczenie pozwoliło autorowi zapoznać się z poszczególnymi aplikacjami do tworzenia filmów deepfake i przełożyło się na wnioski opisane pod koniec rozdziału. Pozwoliło to również na weryfikację hipotezy stawianej w rozdziale, mówiącej, że nagrania deepfake powstają poprzez wykorzystanie oprogramowania do zmiany obrazu lub dźwięku w oryginalnym nagraniu, tak, aby wyglądało to, jakby ktoś inny mówił lub wyglądał inaczej niż w rzeczywistości.

W rozdziale trzecim wykazane zostało, iż obecnie najbardziej dominującą tematyką wykorzystującą deepfake jest pornografia oraz satyra. Zwrócono jednak uwagę, iż rośnie ilość wykorzystania deepfake w tworzeniu nagrań dezinformacyjnych, dyskredytujących daną osobę lub wykorzystywanych w oszustwach.

W toku badań określone zostały parametry sprzętu niezbędnego do tworzenia nagrań deepfake dobrej jakości. Wskazano, iż wymagania te wraz z rozwojem technologii ulegają ciągłym zmianom, a coraz więcej usług deepfake dostępnych jest w Internecie. Ich zakup nie wymaga wiedzy technicznej, a wraz z czasem ich dostawcy zapewniają coraz lepszą jakość. Wnioskowano, iż do tworzenia wysokiej jakości nagrań nadal niezbędna jest dobra wiedza z obszarów informatyki i grafiki komputerowej.

Zwieńczeniem dotychczasowych działań podjętych w drugim i trzecim rozdziale, było utworzenie eksperymentu laboratoryjnego. Analiza zebranego materiału empirycznego

przeprowadzona została w rozdziałach czwartym, piątym i szóstym. Każdy z nich skoncentrowany był na konkretnym problemie, jaki zidentyfikowany został w kontekście bezpieczeństwa narodowego. Dzięki eksperymentowi udało się odpowiedzieć na trzy pomocnicze pytania badawcze oraz zweryfikować stawiane hipotezy. Ponadto, w rozdziale szóstym zawarte zostało dodatkowe pytanie badawcze, mające nadać pracy utylitarny wymiar.

W rozdziale czwartym przeprowadzona została analiza materiału empirycznego dotyczącego percepcji nagrań deepfake przez osoby badane. Stwierdzono, iż odpowiedzi na każde z pytań dotyczących naturalności nagrań w przypadku filmu deepfake nie są istotnie różne od odpowiedzi dotyczących jednego z prawdziwych filmów, prezentującego średnio znanego influencera. Również w zestawieniu z innym filmem (prezentującym średnio znanego influencera) w trzech pytaniach nie wykryto istotnych statystycznie różnic. Potwierdza to prawdziwość założenia, iż część społeczeństwa ma trudności z wykrywaniem filmów deepfake.

Wykazano również, iż brak umiejętności rozróżniania nagrań deepfake od prawdziwych przełożyć może się na wzrost potencjału działań dezinformacyjnych oraz oszustw finansowych. Wskazano obszary szczególnie podatne na wykorzystanie technologii deepfake.

Wszystkie wyniki analizy statystycznej skorelowane zostały z samooceną zdolności weryfikacji fałszywych nagrań, dzięki czemu możliwe było zweryfikowanie stawianych hipotez oraz oceny odbioru nagrań w poszczególnych grupach.

W piątym rozdziale podjęta została analiza materiału empirycznego dotyczącego wpływu nagrań deepfake na bezpieczeństwo personalne. Poruszone tematy to przede wszystkim oszustwa inwestycyjne prowadzone z wykorzystaniem nagrań deepfake oraz różnica w ich skuteczności względem prawdziwych filmów.

Pomimo tego, iż skuteczność w przekonywaniu do fałszywych inwestycji nagrań deepfake, nie jest na tyle duża co filmów prezentujących prawdziwych influencerów, stwierdzono, iż nadal nagrania deepfake stanowią istotne zagrożenie dla bezpieczeństwa personalnego. Dowiedziono, iż przygotowanie stosownego nagrania jest znaczenie tańsze od zatrudnienia i nagrania fałszywej reklamy ze znaną osobą. Ponadto niewielu influencerów chętnych jest na współpracę z oszustami.

Wykazano, iż umiejętność percepcji zmanipulowanych nagrań wideo może mieć istotny wpływ na wybory i opinie internautów. Respondenci, którzy twierdzili, iż sami nie

daliby się oszukać w takim samym stopniu jak inne osoby, często wykazywali większą podatność i w mniejszym stopniu rozpoznawali nagrania deepfake.

Ponadto stwierdzono, iż dalszy rozwój technologii deepfake może przełożyć się na coraz częstsze wprowadzenie w błąd opinii publicznej i większego wpływu oszustów na wybory oraz decyzje ekonomiczne obywateli. Nagrania deepfake mogą wpływać na decyzje polityczne, prognozy ekonomiczne czy wiadomości o firmach, co może przełożyć się na obniżenie zaufania inwestorów oraz ogólną stabilność gospodarczą.

Dla bezpieczeństwa personalnego najistotniejszym jest jednak fakt, iż materiały utworzone z wykorzystaniem technologii deepfake coraz częściej wykorzystywane są w celu oszustw finansowych, takich jak fałszywe inwestycje, fałszywe nagrania rozmów telefonicznych czy wideokonferencji w celu wyłudzenia pieniędzy lub poufnych informacji handlowych. Przełożyć się to może na wzrost ilości oszukanych osób oraz wymaga od państwa oraz instytucji finansowych przeciwdziałania tym scenariuszom.

Pierwsza część szóstego rozdziału ukazuje wskaźniki psychologiczne, których wzmocnienie lub osłabienie w społeczeństwie przekłada się na wzrost podatności na manipulację z wykorzystaniem nagrań deepfake. Interpretacja wyników wymaga uwzględnienia zarówno postępów technologicznych, jak i zmian społecznych.

Wykazano, że osoby impulsywne wierzą, iż inni ludzie będą podejmować decyzje finansowe na podstawie treści marketingowych w formie filmów, niezależnie od tego, czy zostały one stworzone z wykorzystaniem technologii deepfake, czy nie, oraz niezależnie od tego, czy przedstawiają znane, czy nieznane osoby. Można zatem wnioskować, że osoby impulsywne najprawdopodobniej projektują swoje własne cechy na innych. Ponadto, impulsywne jednostki są bardziej skłonne do zainwestowania swoich środków po obejrzeniu fałszywego nagrania, co sugeruje ich większą podatność na manipulację treściami wideo.

Stwierdzono, iż osoby, dla których istotna jest sprawiedliwość (MFQ sprawiedliwość – oszustwo) są bardziej skłonne do negatywnej oceny nagrania i osoby przedstawionej w filmie, niż w przypadku pozostałych osób. Są one również zdolniejsze w kwestionowaniu wiarygodności informacji zawartych w nagraniach. Ponadto wykazano, iż osoby o wyższym szacunku dla władzy i autorytetów (MFQ autorytet – uległość) są bardziej podatne na wpływ autorytetów oraz skłonne do zaufania informacjom przez nie przekazywanym. Dalszą obserwacją było, iż osoby przykładające wagę do czystości i świętości (MFQ Świętość/Upodlenie) bardziej radykalnie oceniają wiarygodność influencerów pod kątem ich zewnętrznej prezencji.

Na podstawie korelacji odpowiedzi z moderatorem Potrzeby Poznawczego Domknięcia stwierdzono, iż wybrane cechy nagrań deepfake wprowadzających odbiorcę w błąd, wywołują u osób o wysokiej nietolerancji wieloznaczności silniejsze reakcje, przez co osoby te mogą być bardziej skłonne do wyciągania szybkich wniosków i formułowania opinii na podstawie niepełnych lub niejednoznacznych informacji. Przekłada się to na zmniejszoną odporność tych osób na nagrania deepfake. Ponadto osoby te, poprzez niejednoznaczność nagrań, mogą mieć również wyolbrzymione postrzeganie ich wpływu na negatywny odbiór wizerunku aktorów przez inne osoby. Wykazano również zależność, zgodnie z którą osoby o wysokiej nietolerancji wieloznaczności są bardziej skłonne postrzegać filmy deepfake jako bardziej manipulacyjne oraz uznawać ich wpływ na decyzje inwestycyjne innych.

Pozostałe z zastosowanych w badaniu moderatorów psychologicznych – moderatory samooceny, depresji i lęku – nie wykazały innych zależności. Występujące pojedyncze korelacje są zbyt słabe lub przypadkowe, by mówić o wykazanym efekcie.

W drugiej części rozdziału szóstego szeroko omówiono obszary najbardziej narażone na konsekwencje internetowej dezinformacji. Nadmieniono, że manipulacyjne filmy są wykorzystywane nie tylko w celu kradzieży tożsamości, ale i oszustw finansowych, podważania zaufania do instytucji czy osobistości publicznych, a nawet mogą stanowić bezpośrednie zagrożenie dla bezpieczeństwa państwa. W miarę upowszechniania się technologii deepfake istnieje realne ryzyko, że stanie się ona narzędziem o globalnym zasięgu, stosowanym przez osoby o różnych motywacjach, od przestępczych po polityczne.

Opisano również dziedziny działań, których wdrożenie może przeciwdziałać analizowanym w ramach pracy procesom. Rozwój algorytmów detekcji deepfake, edukacja społeczeństwa na temat zagrożeń związanych z fałszywymi nagraniami oraz tworzenie odpowiednich ram prawnych to tylko niektóre z czynności, które mogą przyczynić się do zwiększenia bezpieczeństwa w przestrzeni cyfrowej. W tym kontekście kluczową rolę odgrywają nie tylko instytucje państwowe, ale również media, organizacje pozarządowe oraz platformy internetowe, które powinny podejmować aktywne działania mające na celu zwalczanie dezinformacji i fałszywych treści. Podejście to wymaga wieloaspektowego podejścia, obejmującego zarówno działalność społeczną, technologiczne innowacje, jak i współpracę międzysektorową.

W świetle analizy przeprowadzonej w pracy doktorskiej można przyjąć, że hipoteza badawcza została potwierdzona, a technologia deepfake faktycznie stanowi zagrożenie dla

bezpieczeństwa narodowego poprzez generowanie fałszywych nagrań mogących wpływać na decyzje osób je oglądających.

Analiza przeprowadzona w ramach niniejszej pracy doktorskiej pozwala dostrzec, że obywatele nie zawsze są w stanie rozróżnić fałszywe materiały audiowizualne od prawdziwych, a percepcja zmanipulowanych nagrań wideo może mieć istotny wpływ na zaufanie do autorytetów czy decyzje inwestycyjne, co z kolei może prowadzić do nieprzewidywalnych konsekwencji dla społeczeństwa i gospodarki.

Ponieważ technologia deepfake jest niezwykle dynamicznie rozwijającą się, wiąże się to z poważnymi konsekwencjami dla bezpieczeństwa narodowego. Z jednej strony narzędzia do tworzenia nagrań deepfake stają się coraz bardziej dostępne i łatwe w użyciu, co prowadzi do ich powszechniejszego wykorzystania. Z drugiej zaś, rozwój sztucznej inteligencji, w tym przyszłe wersje modeli takich jak GPT-5, może jeszcze bardziej ułatwić tworzenie manipulacyjnych treści wideo.

Tylko poprzez kompleksowe podejście, które łączy działania technologiczne, społeczne i polityczne, można skutecznie przeciwdziałać zagrożeniom związanym z deepfake i zapewnić bezpieczeństwo narodowe oraz osobiste w erze cyfrowej. Jak wykazano w pracy, konieczna jest również wielopoziomowa współpraca. Określona została przede wszystkim na poziomie rządowym, jednak w pracy zaznaczono również duży wpływ na zwalczenie dezinformacji deepfake platform społecznościowych oraz niezależnych organizacji badawczych.

Niniejsza praca stanowi wkład w dyskusję na temat wyzwań związanych z rosnącą rolą technologii deepfake w przestrzeni medialnej i społecznej. Jej wyniki oraz wnioski mogą stanowić cenne wsparcie dla decydentów politycznych, praktyków mediów oraz specjalistów zajmujących się bezpieczeństwem w podejmowaniu działań mających na celu skuteczną ochronę społeczeństwa przed negatywnymi skutkami manipulacji przy użyciu tej technologii.

Bibliografia

Monografie i opracowania zwarte oraz artykuły naukowe

1. Abhishek G. B. K., Achyuth N. S., Dhananjay S., Mrudula P. B., "Determination of fake news using blockchain and IBM Watson", 2020.
2. Adamic L., Glance N., "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog", LinkKDD '05 Proceedings of the 3rd International Workshop on Link Discovery, 2005.
3. Aldwairi M., Alwahedi A., "Detecting Fake News in Social Media Networks", Procedia Computer Science, 141, 2018.
4. Allcott H., Gentzkow M., "Social media and fake news in the 2016 election", Journal of Economic Perspectives 31(2), 2017.
5. Anderson K. E., "Getting acquainted with social networks and apps: combating fake news on social media", Library HiTech News, 35(3), 2018.
6. Araujo T., Neijens P., „Unobtrusive Measures for Media Research", [w:] J. Van den Bulck (red.), "The International Encyclopedia of Media Psychology" (Vol. 3). (The Wiley Blackwell-ICA International Encyclopedias of Communication). Wiley Blackwell, 2021.
7. Babbie E., „Podstawy badań społecznych", Wydawnictwo Naukowe PWN, Warszawa 2013.
8. Babbie E., „Badania społeczne w praktyce", Warszawa: Wydawnictwo Naukowe PWN, 2004.
9. Bakshy E., Messing S., Adamic L. A., "Exposure to Ideologically Diverse News and Opinion on Facebook", Science 348 (6239), 2015.
10. Baran M. i inni, „Cybernauci – diagnoza wiedzy, umiejętności i kompetencji dzieci i młodzieży, rodziców i opiekunów oraz nauczycieli w zakresie bezpiecznego korzystania z Internetu. Raport podsumowujący badanie ex-ante" Warszawa 2016.
11. Basu S., "The conservatism principle and the asymmetric timeliness of earnings", Journal of Accounting and Economics, volume 24, 1997.
12. Bates S., "Revenge porn and mental health: a qualitative analysis of the mental health effects of revenge porn of female survivors", Feminist Criminology, 12(1), 2017.

13. Bednarowska Z., „Desk research – wykorzystanie potencjału danych zastanych w prowadzeniu badań marketingowych i społecznych", Uniwersytet Jagielloński w Krakowie „Marketing i rynek", 7/2015.
14. Berinsky A. J., "Rumors and health care reform: Experiments in political misinformation", British Journal of Political Science, 47(2), 2017.
15. Boehm L. E., 1994. "The validity effect: a search for mediating variables", Personality and Social Psychology Bulletin 20, 3.
16. Borges L., Martins B., Calado P., "Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News", Journal of Data and Information Quality, 11(3): Article No. 14, 2019.
17. Boxell L., Gentzkow M., Shapiro J. M., "Greater Internet use Is Not Associated with Faster Growth in Political Polarization among US Demographic Groups", Proceedings of the National Academy of Sciences 114 (40), 2017.
18. Britt M. A., Rouet J. F., Blaum D., Millis K., "A Reasoned Approach to Dealing with Fake News. Policy Insights from the Behavioral and Brain Sciences", 6(1), 2019.
19. Brucato B., "Policing made visible: Mobile technologies and the importance of point of View", Surveillance & society, 13(3/4), 2015.
20. Brunese L., Martinelli F., Mercaldo F., Santone A., „Machine learning for coronavirus covid-19 detection from chest x-rays", Procedia Computer Science, Volume 176, 2020.
21. Brzeziński J. M., "Metodologia badań psychologicznych", Warszawa 2019.
22. Brzeziński M., „O zagrożeniach codziennych, nadzwyczajnych i sytuacjach kryzysowych z perspektywy systemowej", [w:] S. Sulowski, M. Brzeziński „Trzy wymiary współczesnego bezpieczeństwa" 2014.
23. Bubnovskaia O. V., Leonidova V. V., Lysova A. V., „Security or Safety: Quantitative and Comparative Analysis of Usage in Research Works Published in 2004–2019", Behavioral Sciences 2019.
24. Bullock J. G., (2009). "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion", The Journal of Politics 71 (3).
25. Buzan B., Wæver O., de Wilde J., "Security: a New Framework for Analysis", CO: Lynne Rienner, Londyn 1998.
26. Chawla R., 2019. "Deepfakes: How a pervert shook the world", International Journal of Advance Research and Development, 4(6).

27. Chow S. L., “Experimentation in psychology–rationale, concepts, and issues”, [w:] “Methods in Psychological Research – Encyclopedia of Life Support Systems” (EOLSS), Eolss Publishers, Oxford, UK, 2002.
28. Ciekanowski Z., „Rodzaje i źródła zagrożeń bezpieczeństwa”, [w:] „Bezpieczeństwo i Technika Pożarnicza”, nr. 1, Warszawa 2010.
29. Cieślarczyk M., Chojnacki Z., „Techniki i narzędzia badawcze stosowane w pracach magisterskich i doktorskich”, [w:] „Metody, techniki i narzędzia badawcze oraz elementy statystyki stosowane w pracach magisterskich i doktorskich”, M. Cieślarczyk (red.), Warszawa 2006.
30. Cieślarczyk M., Chojnacki Z., Mróz M., Sirko S., „Metody techniki i narzędzia badawcze oraz elementy statystyki stosowane w pracach magisterskich i doktorskich”, Warszawa 2006.
31. Citron D. K., Franks M. A., “Criminalizing Revenge Porn”, Wake Forest Law Review, Vol. 49, 2014, p. 345+, u of Maryland Legal Studies Research Paper No. 2014-1.
32. Czaputowicz J., „Czy interdyscyplinarność jest właściwym kierunkiem rozwoju stosunków międzynarodowych w Polsce?”, [w:] A. Gałganek, E. Haliżak, M. Pietraś (red.), Wieloi interdyscyplinarność nauki o stosunkach międzynarodowych, Polskie Towarzystwo Stosunków Międzynarodowych, Wydawnictwo Rambler, Warszawa 2012.
33. Czupryński A., „Metoda naukowa”, [w:] Nauki o bezpieczeństwie. Wybrane problemy badań.”, red. A. Czupryński, B. Wiśniecki, J. Zboina, Józefów 2017.
34. Dąbrowska I., „Deepfake – nowy wymiar internetowej manipulacji”, Zarządzanie Mediami. 8, 2020.
35. Day C., “The Future of Misinformation”, Computing in Science & Engineering, 21(1), 2019.
36. De keersmaecker J., Roets A., “‘Fake news’: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions”, Intelligence, 65, 2017.
37. Delfino R. A., “Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act”, 88 Fordham L. Rev. 887, 2019.
38. Dunning D., Griffin D. W., Milojkovic J. D., Ross L., “The overconfidence effect in social prediction”, Journal of Personality and Social Psychology, 58(4), 1990.

39. Durbin J., „Incomplete blocks in ranking experiments", British Journal of Statistical Psychology, 4, 1951.
40. Farrell H., Schneier B., "Common Knowledge Attacks on Democracy", 2018.
41. Faul F., Erdfelder E., Lang A.-G., & Buchner A. "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences", Behavior Research Methods, 39, 2007.
42. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses". Behavior Research Methods, 41, 2009.
43. Fido D., Rao J., Harper C. A., "Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography", Judgements of Deepfake Media Production, 2020.
44. Figueira A., Oliveira L., "The current state of fake news: challenges and opportunities", Procedia Computer Science, 121, 2017.
45. Fish J.M., McCraw S.J., Reddish Ch.J., "Fighting in the gray zone: a strategy to close the preemption gap, US Army War College", Strategic Studies Institute, 2004.
46. Flaxman S., Goel S., Rao J. M., "Filter Bubbles, Echo Chambers, and Online News Consumption", Public Opinion Quarterly 80 (1), 2016.
47. Fletcher J., "Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance", Theatre Journal, 70(4): 455–471. Project MUSE, 2018,
48. Frei D., "Grundfragen der Weltpolitik", Stuttgart 1977. Cyt. za: M. Adamczyk, „Teoretyczne wprowadzenie do badań nad bezpieczeństwem", [w:] M. Debita, M. Adamczyk (red.), „Polska – Europa – Świat. Wczoraj i dziś", Poznań 2017.
49. Frenda S. J., Knowles E. D., Saletan W., Loftus E. F., "False memories of fabricated political events. Journal of Experimental Social Psychology", 49(2), 2013.
50. Frijda N. F., (1986). "The emotions", Cambridge University Press.
51. Gaines B. J., Kuklinski J. H., Quirk P. J., Peyton B., Verkuilen J., "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq", The Journal of Politics 69 (4), 2007.
52. Giedymin J., „Problemy, założenia, rozstrzygnięcia: studia nad logicznymi podstawami nauk społecznych", Poznań 1964.

53. Goel S., Anderson A., Hofman J., Watts D. J., “The structural virality of online diffusion”, Management Science, 62(1), 2015.
54. Grabe M. E., Bucy E. P., „Image bite politics: News and the visual framing of elections”, Oxford University Press, 2009.
55. Graber D. A., “Seeing is remembering: How visuals contribute to learning from television news”, Journal of Communication, 40(3), 1990.
56. Graham J., Nosek B. A., Haidt J., Iyer R., Spassena K., & Ditto P. H., „Moral Foundations Questionnaire (MFQ)” [Database record]. APA PsycTests, 2011.
57. Granot Y., Balcetis E., Feigenson N., Tyler T., “In the eyes of the law: Perception versus reality in appraisals of video evidence”, Psychology, Public Policy, and Law, 24(1), 93–104.”, 2018
58. Guess A., Lyons B., Nyhan B., Reifler J., (2017). “Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Congenial Political News is Less Prevalent than You Think”, Knight Foundation report.
59. Harris D., (2019). “Deepfakes: False pornography is here and the law cannot protect you”, Duke Law & Technology Review,17.
60. Hasan H. R., Salah K., 2019. “Combating Deepfake Videos Using Blockchain and Smart Contracts”, IEEE Access, 7.
61. Hofferth L. (2005). Secondary Data Analysis in Family Research. Journal of Marriage and Family, 67(4).
62. Ismael A.M., Şengür A., „Deep learning approaches for COVID-19 detection based on chest X-ray images”, Expert Systems with Applications, Volume 164, 2021.
63. Jakubczak R., „Obrona narodowa w tworzeniu bezpieczeństwa III RP”, Dom Wydawniczy BELLONA, Warszawa 2003.
64. Jamieson K. H., Cappella J. N., “Echo chamber: Rush Limbaugh and the conservative media establishment”, Oxford University Press; The Spreading of Misinformation Online, 2008.
65. Jemioło T., Dawidczyk A., „Wprowadzenie do metodologii badań bezpieczeństwa”, Warszawa 2008.
66. Johnson M. K., Raye C. R., “Reality monitoring”, Psychological review, 88, 1, 1981.
67. Johnston M.P., „Secondary Data Analysis: a Method of which the Time Has Come”, [w:] “Qualitative and Quantitative Methods in Libraries (QQML)” 2014.

68. Kacała T., „Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa", Przegląd Prawa Konstytucyjnego, nr 2/2015.
69. Kahan D. M., (2013). "Ideology, Motivated Reasoning, and Cognitive Reflection", Judgment and Decision Making 8, no. 4.
70. Kahneman D., Tversky A., (2013). "Prospect theory: An analysis of decision under risk. In Handbook of the fundamentals of financial decisionmaking: Part I", World Scientific.
71. Kazimierczak D., „Walka informacyjna we współczesnych konfliktach i jej społeczne konsekwencje", Studia de Securitate et Educatione Civili 7, 2017.
72. Kietzmann T. C., Geuter S., König P. (2011). "Overt visual attention as a causal factor of perceptual awareness". PloS one, 6(7).
73. Klepka R., „Manipulacja medialna" [w:] O. Wasiuta, R. Klepka, R. Kopeć (red.), „Vademecum bezpieczeństwa", Libron, Kraków 2018.
74. Kmiecik P., „Bezpieczeństwo informacyjne Rzeczypospolitej w dobie Fake News – przykłady wykorzystania mediów cyfrowych w szerzeniu dezinformacji", Bezpieczeństwo Obronność Socjologia, 11/12 2019.
75. Korshunov P. & Marcel S. (2020). "Deepfake detection: humans vs. machines".
76. Korzeniowski L. F., "Podstawy nauk o bezpieczeństwie. Wydanie 2" 2017.
77. Kossowska M., „Różnice indywidualne w potrzebie poznawczego domknięcia", Przegląd Psychologiczny, 46, 2003.
78. Kossowska M., Hanusz K., Trejtowicz M., „Skrócona wersja Skali Potrzeby Poznawczego Domknięcia. Dobór pozycji i walidacja skali", [w: Psychologia Społeczna 2012 tom 7 1 (20)].
79. Kotowicz W., „Bezpieczeństwo narodowe", [w:] A. Żukowski (red.), M. Hartliński (red.), W. T. Modzelewski (red.), J. Więcławski (red.), „Podstawowe kategorie bezpieczeństwa narodowego", Olsztyn 2015.
80. Koziej S., „Bezpieczeństwo: istota, podstawowe kategorie i historyczna ewolucja", [w:] Bezpieczeństwo narodowe 18 (2), 2011.
81. Kroenke K., Spitzer R. L., Williams J. B., „The PHQ-9: validity of a brief depression severity measure", 2001 J Gen Intern Med. 16(9).
82. Kruger J., Dunning D., „Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments", „Journal of Personality and Social Psychology", 77 (6), 1999.

83. Kruskal W., Wallis W., „Use of ranks in one-criterion variance analysis", Journal of the American Statistical Association. 1952, 47 (260).
84. Kuran T., Sunstein C. R., (1999). "Availability cascades and risk regulation", Stanford Law Review.
85. Laguna M., Lachowicz-Tabaczek K., Dzwonkowska I., „Skala Samooceny SES Morrisa Rosenberga – polska adaptacja metody", Psychologia Społeczna 2, 2007.
86. Leibenstein H., 1950. "Bandwagon, snob, and Veblen effects in the theory of consumers' demand", The quarterly journal of economics 64, 2.
87. Lem S., "Kongres futurologiczny", 1983.
88. Lewandowsky S., Ecker U., Seifert C. M., Schwarz N., Cook J., (2012). "Misinformation and its Correction: Continued Influence and Successful Debiasing", Psychological Science in the Public Interest 13 (3).
89. Lin H., Kerr J., (2018). "On Cyber-Enabled Information/Influence Warfare and Manipulation", Oxford University Press: 2018 forthcoming.
90. Lipski S., „Bezpieczeństwo narodowe – wybrane zagadnienia terminologiczne", [w:] T. Jemioło, K. Rajchel, „Bezpieczeństwo narodowe i zarządzanie kryzysowe w Polsce w XXI wieku – wyzwania i dylematy: praca zbiorowa", Warszawa 2008, Cyt. Za: W. Kotowicz, „Bezpieczeństwo narodowe", [w:] A. Żukowski (red.), M. Hartliński (red.), W. T. Modzelewski (red.), J. Więcławski (red.), „Podstawowe kategorie bezpieczeństwa narodowego", Olsztyn 2015.
91. Łobocki M., „Metody badań pedagogicznych", Warszawa 1982.
92. MacLeod C., Mathews A., Tata P., (1986). "Attentional bias in emotional disorders", Journal of abnormal psychology 95, 1.
93. Maras M. H., Alexandrou A., 2019. "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos", International Journal of Evidence & Proof, 23(3).
94. Maslow A., „Motywacja i osobowość", Warszawa 1990.
95. Mider D., Marcinkowska A., „Analiza danych ilościowych dla politologów. Praktyczne wprowadzenie z wykorzystaniem programu GNU PSPP", ACAD, Warszawa 2013.
96. Mider D., „Polacy wobec przemocy politycznej", Warszawa, 2017.

97. Motylińska P. „Manipulacja informacją” [w:] O. Wasiuta, R. Klepka, R. Kopeć (red.), „Vademecum bezpieczeństwa”, Libron, Kraków 2018.
98. Mroczka K., “Fake news as a new category of threat to the system of economic security of the state in the era of epidemic crisis”, Przegląd Bezpieczeństwa Wewnętrznego nr. 26 (14), 2022.
99. Neal R. D., Lawlor D. A., Allgar V., “Missed appointments in general practice: retrospective data analysis from four practices”, [w:] “British Journal of General Practice”, 2001.
100. Neudert L. M., “Future Elections May Be Swayed by Intelligent, Weaponized Chatbots”, MIT Technology Review, 2018.
101. Newman E. J., Garry M., Unkelbach C., Bernstein D. M., Lindsay D., Nash R. A., (2015). “Truthiness and falsiness of trivia claims depend on judgmental contexts”, Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5).
102. Nickerson R. S., (1998). “Confirmation bias: a ubiquitous phenomenon in many guises”, Review of General Psychology, 2(2).
103. Nowak S. (red.), „Metody badań socjologicznych. Wybór tekstów”, Warszawa 1965.
104. Nowak S., „Metodologia badań społecznych”, Warszawa 2010.
105. Nowiński M., „Pojęcie bezpieczeństwo narodowe w prawie europejskim i międzynarodowym w kontekście uprawnień służb specjalnych”, [w:] „Uprawnienia Służb Specjalnych z Perspektywy Współczesnych Zagrożeń Bezpieczeństwa Narodowego. Wybrane Zagadnienia”, Warszawa 2017.
106. Nyhan B., Reifler J., (2010). When Corrections Fail: The Persistence of Political Misperceptions. Political Behavior 32(2).
107. Oates S., (2018). “When Media Worlds Collide: Using Media Model Theory to Understand How Russia Spreads Disinformation in the United States”.
108. Pancer E., Poole M., (2016). “The popularity and virality of political social media: Hashtags, mentions, and links predict likes and retweets of 2016 US presidential nominees’ tweets”, Social Influence, 11(4).
109. Pawłuszko T., „Wstęp do metodologii badań politologicznych”, Częstochowa 2013.

110. Pennycook G., Cannon T. D., Rand D. G., (2018). “Prior Exposure Increases Perceived Accuracy of Fake News”, Journal of Experimental Psychology: General 147, no. 12.
111. Pennycook G., Rand D. G., (2017). “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings”, Working paper.
112. Petrov I. & Gao D. & Chervoniy N. & Liu K. & Marangonda S. & Umé C. & Jiang J. & RP L. & Zhang S. & Wu P. & Zhang W. “DeepFaceLab: a simple, flexible and extensible face swapping framework”, 2020.
113. Pieszko – Sroka A., „Czy zeznania są wiarygodne? Poszukiwanie metody ich oceny i rola psychologa w tym procesie”, Przegląd Bezpieczeństwa Wewnętrznego 5/2011.
114. Pieter J., „Ogólna metodologia pracy naukowej”, Wrocław 1967.
115. Piwowarski J. A., „Metodologiczne i badawcze założenia pracy dyplomowej z dyscypliny nauk o bezpieczeństwie – przykład”, [w:] Security, Economy & Law Nr 4/2019 (XXV).
116. Pogue D., (2017), “How to stamp out fake news”, Scientific American, 20316(2).
117. Pokruszyński W., „Teoretyczne aspekty bezpieczeństwa”, Józefów 2010, Cyt. Za: A. Wawrzusiszyn (red.), „Praca dyplomowa z bezpieczeństwa – wprowadzenie do badań”, Warszawa 2016.
118. Prior M., (2013). “Visual political knowledge: a different road to competence?”, Journal of Politics, 76(1).
119. Pronin E., Kruger J., Savitsky K., Lee R.K. (2001). “You Don’t Know Me, But i Know You: The Illusion of Asymmetric Insight”, Journal of Personality and Social Psychology – PSP. 81.
120. Qayyum A., Qadir J., Janjua M. U., Sher F., 2019. “Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News”, IT Professional, 21(4).
121. Schwarz N., Sanna L. J., Skurnik I., Yoon C. (2007). „Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns”, Advances in Experimental Social Psychology, 39.

122. Sekściński A., „Bezpieczeństwo wewnętrzne w ujęciu teoretycznym. Geneza i współczesne rozumienie w naukach politycznych", e-Politikon 2013, nr 6.
123. Sienkiewicz P., „Metody badań nad bezpieczeństwem i obronnością", Warszawa 2010.
124. Spitzer R. L., Kroenke K., Williams J. B. W., Löwe B., „A brief measure for assessing generalized anxiety disorder. Archives of Internal Medicine", 166(10), 2006.
125. Spitzer R. L., Kroenke K., Williams J. B. W., 'Patient Health Questionnaire Study Group. Validity and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study", 1999 JAMA;282.
126. Stenberg G., (2006). "Conceptual and perceptual factors in the picture superiority effect", European Journal of Cognitive Psychology, 18(6).
127. Sułek A., „Eksperyment w badaniach społecznych", Warszawa 1979.
128. Sulowski S., „O rozwoju badań i postulacie interdyscyplinarności w naukach o bezpieczeństwie", [w:] Sulowski S. (red.), „Tożsamość nauk o bezpieczeństwie", Toruń 2015.
129. Sütterlin S. & Ask T. & Mägerle S. & Gloeckler S. & Wolf L. & Schray J. & Chandi A. & Bursac T. & Khodabakhsh A. & Knox B. & Canham M. & Lugo R. (2021). "Individual Deep Fake Recognition Skills are Affected by Viewers' Political Orientation, Agreement with Content and Device Used".
130. Szczurek T., Górnikiewicz M., „Social Media Wars – The R-evolution Has Just Begun", Warszawa 2018.
131. Sztumski J., „Wstęp do metod i technik badań społecznych", Katowice 2005.
132. Taber C. S., Lodge M., (2006). Motivated skepticism in the evaluation of political beliefs. American Journal of Political Science 50 (3).
133. Undeutsch U., (1967). Beurteilung der glaubhaftigkeit von aussagen. Handbuch der psychologie11.
134. Vaccari C., Chadwick A., "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News", Social Media + Society, 2020.
135. Volkoff V., Dezinformacja – oręż wojny, Warszawa 1991, Cyt. Za: T. Kacała, „Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa", Przegląd Prawa Konstytucyjnego, nr 2/2015.

136. Volkoff V., Psychosocjotechnika, dezinformacja – oręż wojny, Komorów 1999.
137. Ward A., Ross L., Reed E., Turiel E., Brown T., (1997). "Naive realism in everyday life: Implications for social conflict and misunderstanding". Values and Knowledge.
138. Wardle C., Derakshan H., "Information Disorder Toward an interdisciplinary framework for research and policymaking", Council of Europe report DGI, 2019.
139. Warecki M., Warecki W., „Słowo o manipulacji, czyli krótki podręcznik samoobrony", Poltext, Warszawa 2006.
140. Wasiuta O., Klepka R., Kopeć R. (red.), „Vademecum bezpieczeństwa", Libron, Kraków 2018.
141. Wasiuta O., Klepka R., „Vademecum bezpieczeństwa informacyjnego", 2020, t. 1 A-M.
142. Wasiuta O., Wasiuta S., „Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość", Studia de Securitate 9(3). 2019.
143. Wawrzusiszyn A. (red.), „Praca dyplomowa z bezpieczeństwa – wprowadzenie do badań", Warszawa 2016.
144. Webb E. J., Campbell D. T., Schwartz R. D., Sechrest L., "Unobtrusive measures: Nonreactive research in the social sciences", Chicago, IL: Rand McNally, 1966.
145. Webb E.J., Campbell D.T., Schwartz R.D., Sechrest L., Grove J.B., "Nonreactive Measures in the Social Sciences", Dallas: Houghton Mifflin, 1981.
146. Webster D.M., Kruglanski A. W., „Individual differences in need for cognitive closure", Journal of Personality and Social Psychology, 67, 1994.
147. Westling J., "Deep Fakes: Let's Not Go Off the Deep End", 2019.
148. Wiśniewski B. (red.), „Od nauk wojskowych do nauk o bezpieczeństwie", Szczytno 2014.
149. Witten I. B., Knudsen E. I., (2005). "Why seeing is believing: Merging auditory and visual worlds", Neuron, 48(3).
150. Wróblewski R., „Podstawowe pojęcia z dziedziny polityki bezpieczeństwa, strategii i sztuki wojennej", Warszawa 1993.
151. Xu Y., Yang Y., Wang E, Zhuang F., Xiong H., „Detect Professional Malicious User with Metric Learning in Recommender Systems", Journal of Latex Class Files, Vol. 14, No. 8, August 2020.

152. Zaczyński W., “Praca badawcza nauczyciela”, Warszawa 1995.
153. Zięba R., „Pojęcie i istota bezpieczeństwa państwa w stosunkach międzynarodowych”, „Sprawy Międzynarodowe” nr. 10, 1989.
154. Zięba R., „Pojmowanie bezpieczeństwa międzynarodowego w XXI wieku”, [w:] R. Zięba (red.), „Bezpieczeństwo międzynarodowe w XXI wieku”.
155. Znamierowski C. „Szkoła prawa. Rozważania o państwie”, Warszawa 1999.
156. Zuckerman M., DePaulo B. M., Rosenthal R., “Verbal and Nonverbal Communication of Deception”, In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 14, pp. 1-59). New York: Academic Press, 1981.

Źródła internetowe

1. „(10) (PDF) MFQ-PL – Kwestionariusz do pomiaru kodów moralnych”, https://www.researchgate.net/publication/281274953_MFQ-PL_-_Kwestionariusz_do_pomiaru_kodow_moralnych [dostęp: 01.08.2022].
2. Akta opisujące sprawę oszustwa, https://www.documentcloud.org/documents/21085009-hackers-use-deep-voice-tech-in-400k-theft [dostęp: 01.01.2023].
3. Artykuł Alana D. Thompson, komentujący wiadomości dotyczące kolejnych wersji czatu GPT-3, https://lifearchitect.ai/gpt-4/ [dostęp: 15.01.2023].
4. Artykuł na blogu California Coast Line, opisujący działania w sprawie zdjęcia domu Barbary Streisand oraz kontrowersji z tym związanych https://www.californiacoastline.org/news/sjmerc5.html [dostęp:01.05.2024].
5. Artykuł opisujący jeden z przypadków, w którym emerytka straciła wszystkie swoje oszczędności, https://cebrf.knf.gov.pl/komunikaty/artykuly-csirt-knf/362-ostrzezenia/852-66-latka-stracila-prawie-190-tysiecy-zlotych-inwestujac-na-falszywej-platformie [dostęp: 01.01.2023].
6. Artykuł opisujący możliwości wykorzystania deepfake na cenę akcji, https://cointelegraph.com/news/here-s-how-to-quickly-spot-a-deepfake-crypto-scam-cybersecurity-execs [dostęp: 01.01.2023].
7. Artykuł opisujący próby podszywania się pod znanych inwestorów, https://cointelegraph.com/news/sam-bankman-fried-deepfake-attempts-to-scam-investors-impacted-by-ftx [dostęp: 01.01.2023].
8. Artykuł opisujący przypadek chińskiej vlogerki „Your Highness Qiao Biluo”, https://www.bbc.com/news/blogs-trending-49151042 [dostęp: 01.01.2023].

9. Artykuł opisujący reakcję mediów społecznościowych na pojawienie się nagrania, https://edition.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html [dostęp: 01.01.2023].
10. Artykuł opisujący wydarzenia z końca lutego 2022 roku, https://dziendobry.tvn.pl/newsy/problemy-z-wyplacaniem-gotowki-czy-pieniedzy-moze-zabraknac-5617004 [dostęp: 01.01.2023].
11. Artykuł opisujący, jak oszuści utworzyli fałszywe nagranie Elona Muska i zachęcali przy jego pomocy do inwestowania na nieistniejącej giełdzie, https://www.outlookindia.com/business/criminals-use-elon-musk-s-deepfake-video-to-dupe-crypto-investors-crypto-market-rises-news-198403 [dostęp: 01.01.2023].
12. Artykuł Roger Montti dotyczący możliwości dalszego rozwoju GPT-4, https://www.searchenginejournal.com/openai-gpt-4/476759/#close [dostęp: 21.01.2023].
13. Blog twórcy oprogramowania MyFakeApp, https://radek350.wordpress.com/2018/02/17/myfakeapp-fakeapp-alternative/ [dostęp: 01.01.2023].
14. Dodge A., „Using Fake Video Technology to Perpetuate Intimate Partner Abuse – Domestic Violence Advisory", https://www.cpedv.org/sites/main/files/webform/deepfake_domestic_violence_advisory.pdf [dostęp:01.10.2022].
15. Komunikat dotyczący braku planów w zakresie wprowadzania limitów wypłat, https://x.com/uknf/status/1496876307573096452 [dostep: 01.01.2023].
16. Komunikat Narodowego Banku Polskiego zapewniający o bezpiecznym stanie gotówki, https://www.prawo.pl/podatki/wyplaty-gotowki-w-banku-i-z-bankomatu,513700.html [dostęp: 01.01.2023].
17. Komunikat Policji opisujący zgłoszenie utraty oszczędności przez 77-letnią kobietę. https://opolska.policja.gov.pl/op/aktualnosci/115700,77-latka-chciala-zainwestowac-pieniadze-w-Internecie-stracila-270-tysiecy-zlotyc.html [dostęp: 01.01.2024].
18. PQStat - Baza Wiedzy o statystyce, testy nieparametryczne, https://manuals.pqstat.pl/en:statpqpl:porown3grpl:nparpl [dostęp: 01.01.2023].

19. Przykłady fałszywych nagrań deepfake, zachęcających do fałszywej inwestycji dostępne są na stronie Centrum Edukacji dla Bezpieczeństwa Rynku Finansowego, https://cebrf.knf.gov.pl/deepfake [dostęp: 01.02.2024].
20. Pu J., Mangaokar N., „Deepfake Videos in the Wild: Analysis and Detection", 2021, https://arxiv.org/pdf/2103.04263.pdf [dostęp:15.06.2022].
21. Raport CSIRT KNF – Fałszywe Inwestycje, https://cebrf.knf.gov.pl/images/Raporty/Faszywe_inwestycje_2022.pdf [dostęp: 01.01.2023].
22. Spitzer R. L, Williams J.B.W., Kroenke K. oraz współpracownicy z wykorzystaniem grantu oświatowego od firmy Pfizer Inc, https://polpharmadlaciebie.pl/materialy-dla-pacjenta/psychatria/kwestionariusz-phq-9 [dostęp:01.01.2023].
23. Spitzer R.L., Dr Janet B.W. Williams, Dr Kurt Kroenke oraz współpracownicy, z wykorzystaniem grantu oświatowego od firmy Pfizer Inc, https://polpharmadlaciebie.pl/materialy-dla-pacjenta/psychatria/kwestionariusz-gad-7 [dostęp:01.01.2023].
24. Strona główna projektu DISARM Framework https://www.disarm.foundation/framework [dostęp: 30.03.2024].
25. Strona główna Wydziału Bezpieczeństwa, Logistyki i Zarządzania WAT, zalecenie Dziekana WBLiZ, https://wlo.wat.edu.pl/wp-content/uploads/2020/04/wytyczne.pdf [dostęp: 01.12.2022].

Spis ilustracji

Spis tabel

Spis wykresów

Załącznik 1 – Metryczka

Protokół z obliczeń minimalnej próby:

[1] *-– Friday, January 28, 2022 -– 15:42:02*
F tests – ANOVA: Repeated measures, within factors

Analysis:	A priori: Compute required sample size		
Input:	Effect size f	=	0.15
α err prob	=	0.05	
Power (1-β err prob)	=	0.95	
Number of groups	=	1	
Number of measurements	=	6	
Corr among rep measures	=	0.5	
Nonsphericity correction ε	=	1	
Output:	Noncentrality parameter λ	=	20.2500000
Critical F	=	2.2383820	
Numerator df	=	5.0000000	
Denominator df	=	370	
Total sample size	=	75	
Actual power	=	0.9518099	

Statystyki opisowe metryki badanych osób

	zainteresowanie_polityka	liczba_znajomych_media_sp	partia_za_miesiac	partia_ostatnia	godziny_w_Internecie	stan_cywilny	dochody	praca_matka_14lat	praca_ojciec_14lat	sytuacja_zawodowa	wykształcenie	miejsce_zamieszkania	plec	wiek
N	76	77	63	69	79	73	68	74	71	78	79	79	69	73
Missing	6	5	19	13	3	9	14	8	11	4	3	3	13	9
Mean	3	3	6	5	4	3	5	4	5	2	3	2	1	25
Błąd standardowy średniej	0	0	0	0	0	0	0	0	0	0	0	0	0	1
95% CI średniej dolna granica	3	3	5	5	4	3	4	4	4	2	3	2	0	24
95% CI średniej górna granica	3	3	6	6	5	3	5	5	6	3	3	3	1	27
Median	3	3	6	6	4	2	5	4	6	2	3	2	1	25
Mode	2	3	7	7	4	2	8	1	1	1	2	1	1	25
Suma	214	225	355	378	345	226	336	321	347	167	218	193	42	1829
Standard deviation	1	1	1	2	2	2	2	3	3	2	1	2	0	6
Variance	1	1	2	3	3	3	5	9	8	4	1	3	0	38
IQR	2	2	2	2	3	3	4	5	5	1	,	2	3	1
Range	4	5	5	5	7	4	7	9	9	7	4	5	1	45
Minimum	1	1	2	2	1	1	1	1	1	1	1	1	0	18
Maximum	5	6	7	7	8	5	8	10	10	8	5	6	1	63
Skewness	0	0	-1	-1	1	0	0	1	0	2	0	1	0	3
Std. error skewness	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Kurtosis	-1	0	0	-1	0	-2	-1	-1	-1	4	-1	-1	-2	20
Std. error kurtosis	1	1	1	1	1	1	1	1	1	1	1	1	1	1
W Shapiro-Wilka	1	1	1	1	1	1	1	1	1	1	1	1	1	1
wartość p testu Shapiro-Wilka	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001
25trzeci percentyl	2	2	5	5	3	2	3	1	2	1	2	1	0	22
50trzeci percentyl	3	3	6	6	4	2	5	4	6	2	3	2	1	25
75trzeci percentyl	4	4	7	7	6	5	7	6	7	2	4	4	1	27

Preferencje wyborcze badanych osób, gdyby wybory odbyły się za miesiąc

partia_za_miesiac	Liczebności	% całości	% skumulowany
KOMITET WYBORCZY PRAWO i SPRAWIEDLIWOŚĆ	2	3.2 %	3.2 %
KOMITET WYBORCZY SOJUSZ LEWICY DEMOKRATYCZNEJ	6	9.5 %	12.7 %
KOMITET WYBORCZY KONFEDERACJA WOLNOŚĆ i NIEPODLEGŁOŚĆ	3	4.8 %	17.5 %
KOALICYJNY KOMITET WYBORCZY KOALICJA OBYWATELSKA PO. N IPL ZIELONI	19	30.2 %	47.6 %

partia_za_miesiac	Liczebności	% całości	% skumulowany
KOMITET WYBORCZY WYBORCÓW KOALICJA BEZPARTYJNI i SAMORZĄDOWCY	5	7.9 %	55.6 %
Nie głosowałem/łam	28	44.4 %	100.0 %

Preferencje wyborcze badanych osób, głosy oddane w wyborach 2018

partia_ostatnia	Liczebności	% całości	% skumulowany
KOMITET WYBORCZY PRAWO i SPRAWIEDLIWOŚĆ	7	10.1 %	10.1 %
KOMITET WYBORCZY SOJUSZ LEWICY DEMOKRATYCZNEJ	5	7.2 %	17.4 %
KOMITET WYBORCZY KONFEDERACJA WOLNOŚĆ i NIEPODLEGŁOŚĆ	4	5.8 %	23.2 %
KOALICYJNY KOMITET WYBORCZY KOALICJA OBYWATELSKA PO. N IPL ZIELONI	18	26.1 %	49.3 %
KOMITET WYBORCZY WYBORCÓW KOALICJA BEZPARTYJNI i SAMORZĄDOWCY	2	2.9 %	52.2 %
Nie głosowałem/łam	33	47.8 %	100.0 %

Stan cywilny badanych osób

stan_cywilny	Liczebności	% całości	% skumulowany
Pozostający w związku małżeńskim (prawnie uznawanym)	10	13.7 %	13.7 %
W związku partnerskim	33	45.2 %	58.9 %
Żadna z powyższych (NIGDY nie byłem żonaty – nie posiadałem partnerki/ NIGDY nie byłam zamężna – nie posiadałam partnera)	30	41.1 %	100.0 %

Dochody osób badanych

dochody	Liczebności	% całości	% skumulowany
Poniżej 2000	3	4.4 %	4.4 %
2001-3000	5	7.4 %	11.8 %
3001-4000	14	20.6 %	32.4 %
4001-5000	12	17.6 %	50.0 %
5001-6000	7	10.3 %	60.3 %
6001-7000	6	8.8 %	69.1 %
7001-8000	6	8.8 %	77.9 %
Powyżej 8001	15	22.1 %	100.0 %

Zawód rodzica – matka – w wieku 14 lat osoby badanej

praca_matka_14lat	Liczebności	% całości	% skumulowany
Wolne zawody i specjaliści np.: lekarz – nauczyciel – inżynier – artysta – główny księgowy	19	25.7 %	25.7 %
Wyższe stanowiska administracyjne np.: bankowiec – dyrektor w dużej firmie – wysoki urzędnik państwowy – wyższy rangą działacz związkowy	3	4.1 %	29.7 %
Zawody związane z pracą biurową np.: sekretarz prezesa/dyrektora – urzędnik, referent – kierownik biura – pracownik działu księgowości	14	18.9 %	48.6 %
Sprzedaż, handel np.: kierownik działu sprzedaży – właściciel sklepu – sprzedawca – agent ubezpieczeniowy	8	10.8 %	59.5 %
Usługi np.: właściciel restauracji – policjant – kelner – fryzjer – żołnierz zawodowy	6	8.1 %	67.6 %
Wykwalifikowany pracownik fizyczny np.: brygadzista – mechanik samochodowy – drukarz – ślusarz narzędziowy – elektryk	7	9.5 %	77.0 %
Półwykwalifikowany pracownik fizyczny np.: murarz – kierowca autobusu – pracownik w przetwórstwie spożywczym– stolarz – blacharz – piekarz	4	5.4 %	82.4 %
Niewykwalifikowany pracownik fizyczny np.: robotnik – niewykwalifikowany pracownik w fabryce, magazynie – portier	3	4.1 %	86.5 %
Rolnictwo np.: rolnik – robotnik rolny – traktorzysta – rybak	2	2.7 %	89.2 %

Zawód rodzica – matka – w wieku 14 lat osoby badanej

praca_matka_14lat	Liczebności	% całości	% skumulowany
Była bezrobotna	8	10.8 %	100.0 %

Zawód rodzica – ojciec – w wieku 14 lat osoby badanej

praca_ojciec_14lat	Liczebności	% całości	% skumulowany
Wolne zawody i specjaliści np.: lekarz – nauczyciel – inżynier – artysta – główny księgowy	16	22.5 %	22.5 %
Wyższe stanowiska administracyjne np.: bankowiec – dyrektor w dużej firmie – wysoki urzędnik państwowy – wyższy rangą działacz związkowy	3	4.2 %	26.8 %
Zawody związane z pracą biurową np.: sekretarz prezesa/dyrektora – urzędnik, referent – kierownik biura – pracownik działu księgowości	5	7.0 %	33.8 %
Sprzedaż, handel np.: kierownik działu sprzedaży – właściciel sklepu – sprzedawca – agent ubezpieczeniowy	7	9.9 %	43.7 %
Usługi np.: właściciel restauracji – policjant – kelner – fryzjer – żołnierz zawodowy	4	5.6 %	49.3 %
Wykwalifikowany pracownik fizyczny np.: brygadzista – mechanik samochodowy – drukarz – ślusarz narzędziowy – elektryk	11	15.5 %	64.8 %
Półwykwalifikowany pracownik fizyczny np.: murarz – kierowca autobusu – pracownik w przetwórstwie spożywczym– stolarz – blacharz – piekarz	15	21.1 %	85.9 %
Niewykwalifikowany pracownik fizyczny np.: robotnik – niewykwalifikowany pracownik w fabryce, magazynie – portier	3	4.2 %	90.1 %
Rolnictwo np.: rolnik – robotnik rolny – traktorzysta – rybak	3	4.2 %	94.4 %
Był bezrobotny	4	5.6 %	100.0 %

Sytuacja zawodowa badanych osób

sytuacja_zawodowa	Liczebności	% całości	% skumulowany
Wykonuję pracę odpłatną (praca najemna, na własny rachunek, w firmie rodzinnej, we własnym gospodarstwie rolnym) – także w przypadku tymczasowej przerwy w pracy	39	50.0 %	50.0 %
Uczę się w szkole lub na uczelni (nauka nie jest opłacana przez pracodawcę) – także w przypadku wakacyjnej przerwy w nauce	27	34.6 %	84.6 %
Jestem bezrobotny/-a i aktywnie poszukuję pracy	3	3.8 %	88.5 %
Jestem bezrobotny/-a, nie poszukuję aktywnie pracy, ale chciałbym/-abym pracować	1	1.3 %	89.7 %
Jestem na emeryturze, rencie	1	1.3 %	91.0 %
Zajmuję się domem, opiekuję się dziećmi lub innymi osobami	1	1.3 %	92.3 %
(Inna sytuacja)	6	7.7 %	100.0 %

Kierunek studiów badanych osób

kierunek_studiow	Liczebności	% całości	% skumulowany
Stosunki miedzynarodowe	1	2.2 %	2.2 %
Logistyka	1	2.2 %	4.4 %
Nauki o bezpieczeństwie.	1	2.2 %	6.7 %
Obronność Państwa – WAT	1	2.2 %	8.9 %
logistyka	1	2.2 %	11.1 %
socjologia	1	2.2 %	13.3 %
Informatyka	2	4.4 %	17.8 %
Obronność Państwa	2	4.4 %	22.2 %
Bezpieczeństwo wew	1	2.2 %	24.4 %
medycyna	3	6.7 %	31.1 %
Prawo	1	2.2 %	33.3 %
ukończone: mgr rachunkowość i controlling w trakcie: mgr psychologia	1	2.2 %	35.6 %
Pedagogika, studia inżynierskie	1	2.2 %	37.8 %
Ogrodnictwo	1	2.2 %	40.0 %
Kryminalistyka, bezpieczeństwo wewnętrzne	1	2.2 %	42.2 %
WAT OBRONNOŚĆ PAŃSTWA	1	2.2 %	44.4 %

Kierunek studiów badanych osób

	kierunek_studiow	Liczebności	% całości	% skumulowany
	Informatyka, inżynier	1	2.2 %	46.7 %
	Obronność państwa	1	2.2 %	48.9 %
	Logistyka międzynarodowa	1	2.2 %	51.1 %
	Sport	1	2.2 %	53.3 %
	Finanse i rachunkowość	2	4.4 %	57.8 %
	Farmacja	1	2.2 %	60.0 %
	Kierunek lekarski.	1	2.2 %	62.2 %
	Lekarski	1	2.2 %	64.4 %
	Nauczenie biologii	1	2.2 %	66.7 %
cyfrowej	Bezpieczeństwo w gospodarce	1	2.2 %	68.9 %
	Filologia angielska	1	2.2 %	71.1 %
	brak	1	2.2 %	73.3 %
	Ekonomia	1	2.2 %	75.6 %
	Brak	1	2.2 %	77.8 %
	Zarządzanie zasobami ludzkimi	1	2.2 %	80.0 %
	Żaden	1	2.2 %	82.2 %
	Technik Reklamy	1	2.2 %	84.4 %
	Administracja	1	2.2 %	86.7 %
	Inżynier	1	2.2 %	88.9 %
	Socjologia	1	2.2 %	91.1 %
	Lotnictwo i kosmonautyka	1	2.2 %	93.3 %
	nie	1	2.2 %	95.6 %
	Bezpieczeństwo wewnętrzne	1	2.2 %	97.8 %
	ekonomia	1	2.2 %	100.0 %

Wykształcenie badanych osób

wykształcenie	Liczebności	% całości	% skumulowany
Podstawowe	2	2.5 %	2.5 %
Średnie	36	45.6 %	48.1 %
Wyższe niepełne (licencjat)	21	26.6 %	74.7 %
Wyższe pełne (magister)	19	24.1 %	98.7 %
Doktorat	1	1.3 %	100.0 %

Miejsce zamieszkania badanych osób

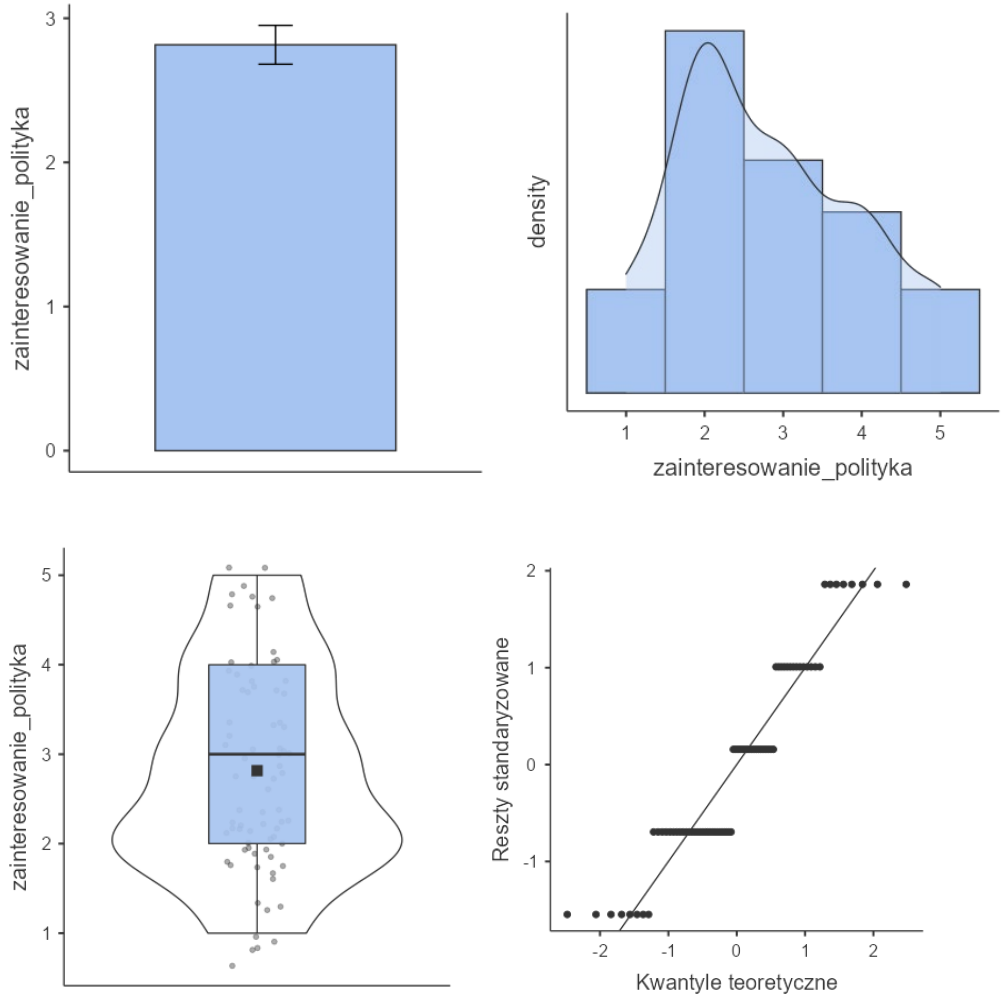
miejsce_zamieszkania	Liczebności	% całości	% skumulowany
Miasto 500 tys. i więcej	37	46.8 %	46.8 %
Miasto 200 – 499 tys.	11	13.9 %	60.8 %
Miasto 100 – 199 tys.	9	11.4 %	72.2 %
Miasto 20 – 99 tys.	7	8.9 %	81.0 %
Miasto poniżej 20 tys.	11	13.9 %	94.9 %
Wieś	4	5.1 %	100.0 %

Płeć badanych osób

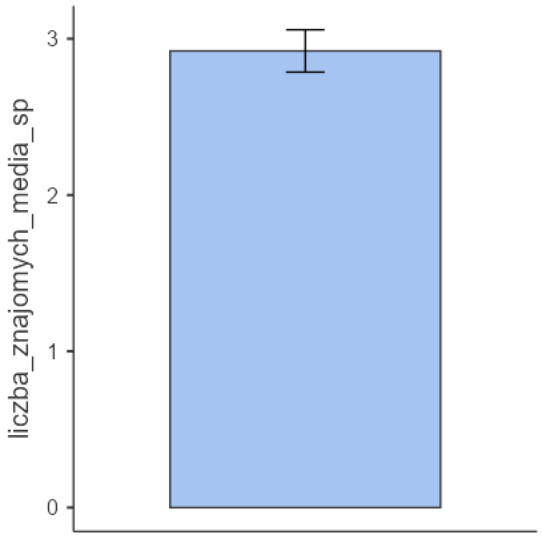
płeć	Liczebności	% całości	% skumulowany
mężczyzna	47	39.1 %	39.1 %
kobieta	72	60.9 %	100.0 %

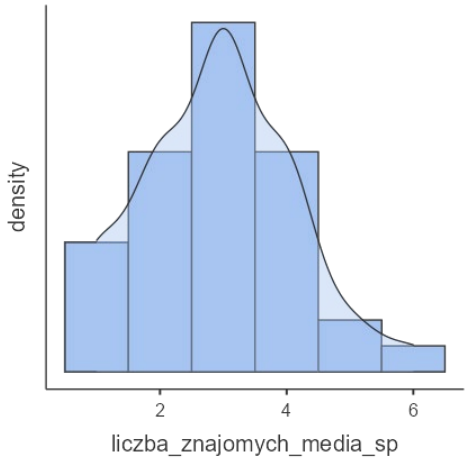
Załącznik 2 – Dodatkowe analizy statystyczne metryczki

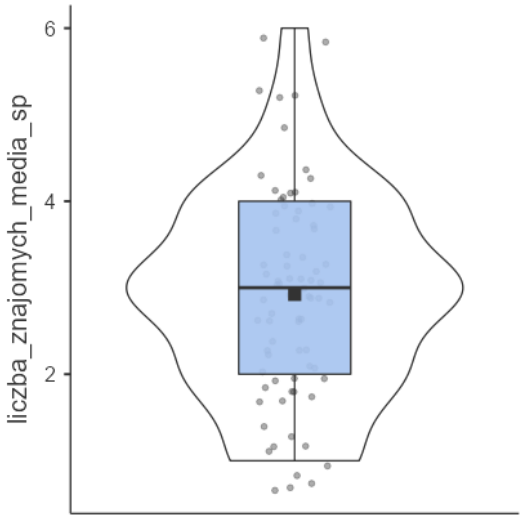
Zainteresowanie polityką wśród badanych osób (1 – w ogóle, 5 – bardzo):

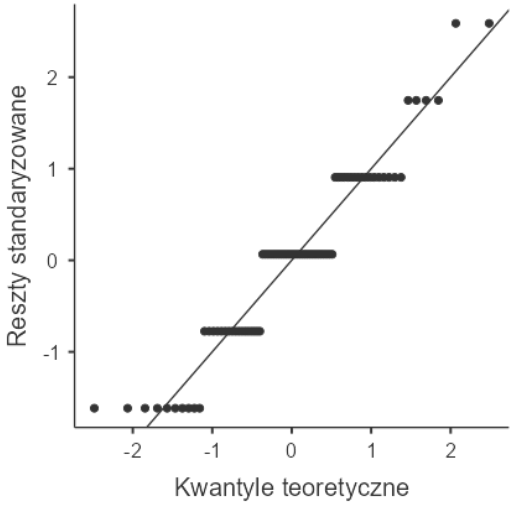


Liczba znajomych w mediach społecznościowych (1 – <100, 2 – 100-200, 3 – 200-300, 4 – 300-400, 5 – 400-500, 6 – 600<):

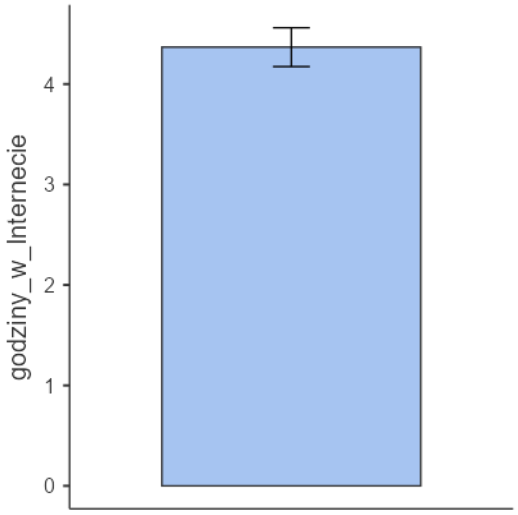


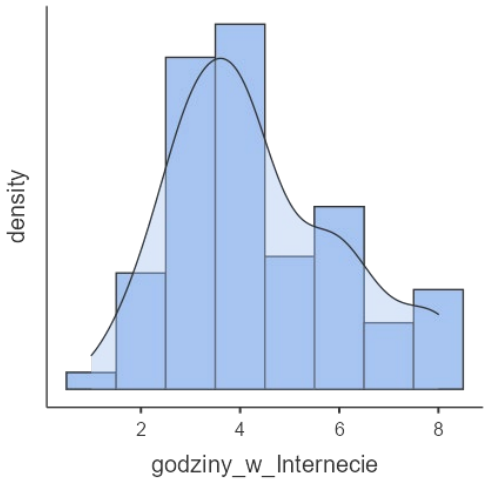


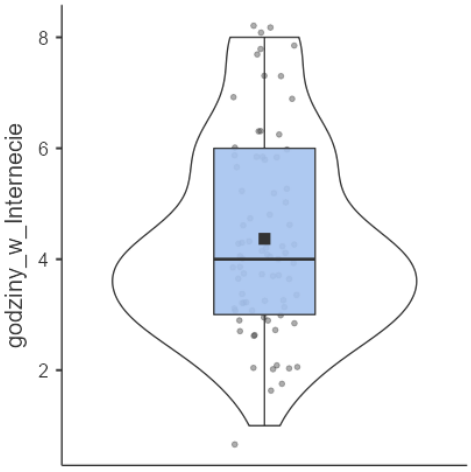


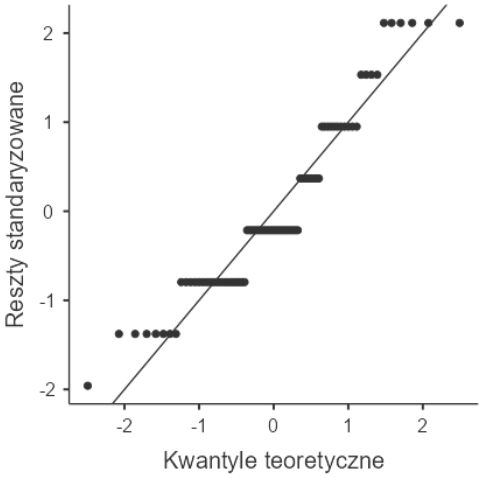


Deklarowana liczba godzin spędzana dziennie w Internecie:

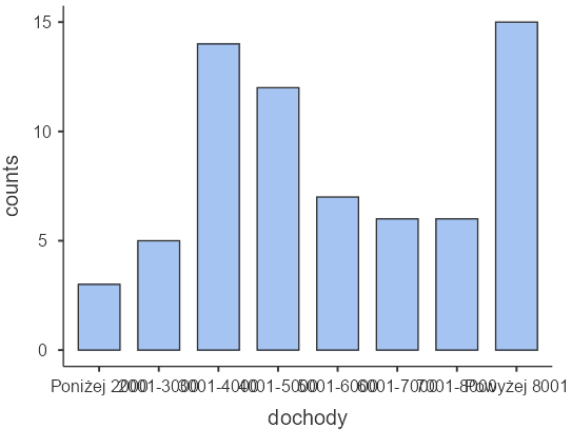


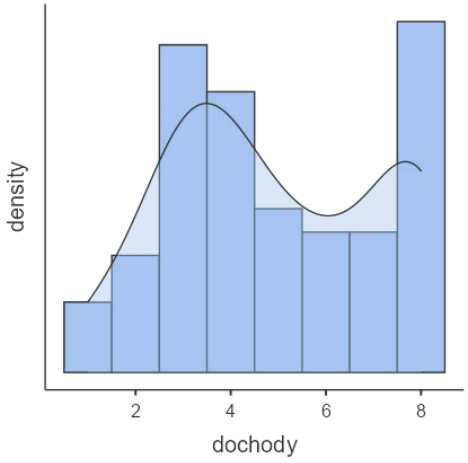


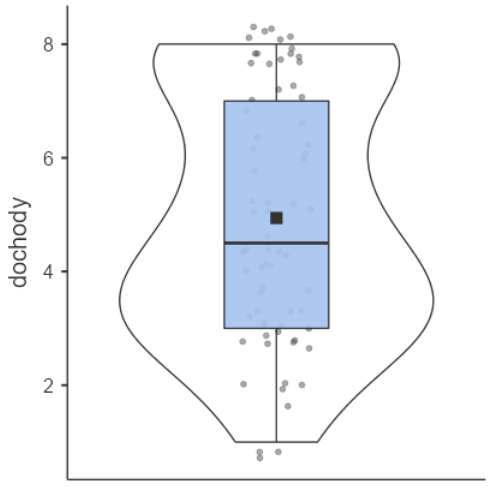


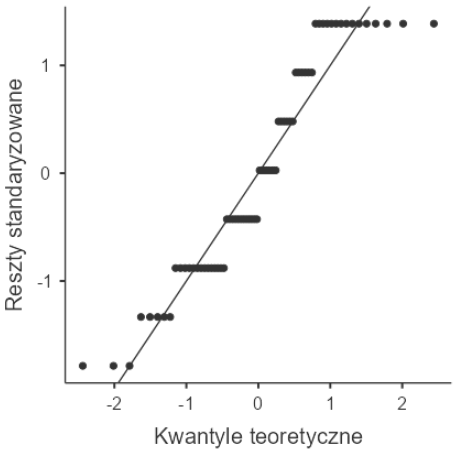


Deklarowane dochody:

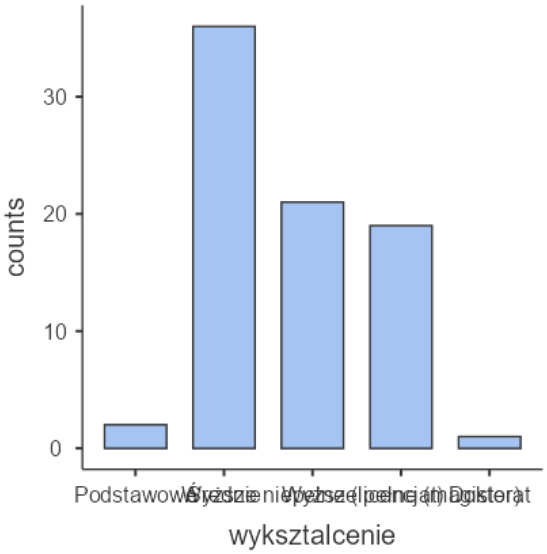


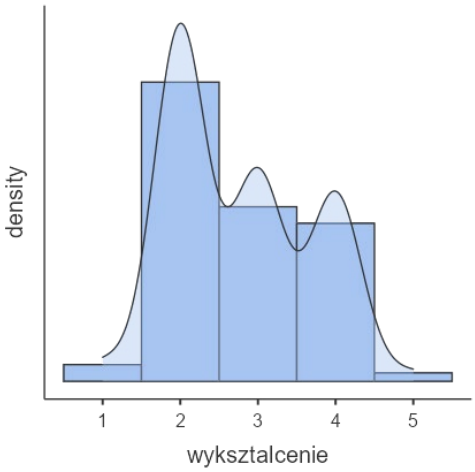


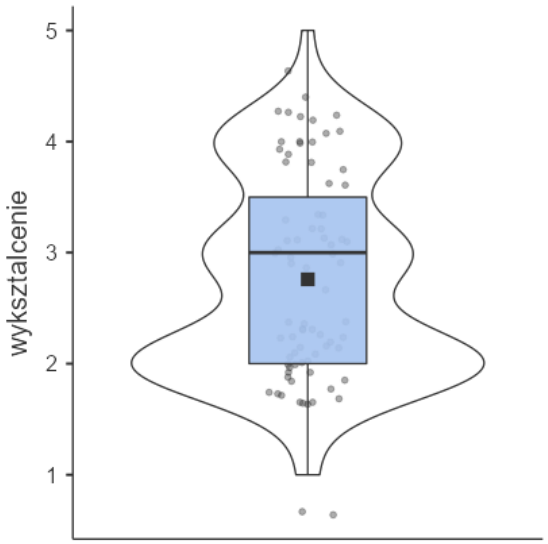


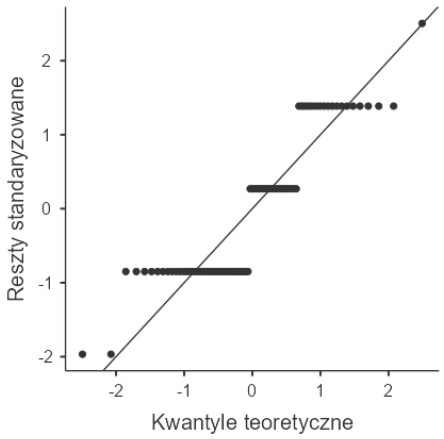


Deklarowane wyksztalcenie:

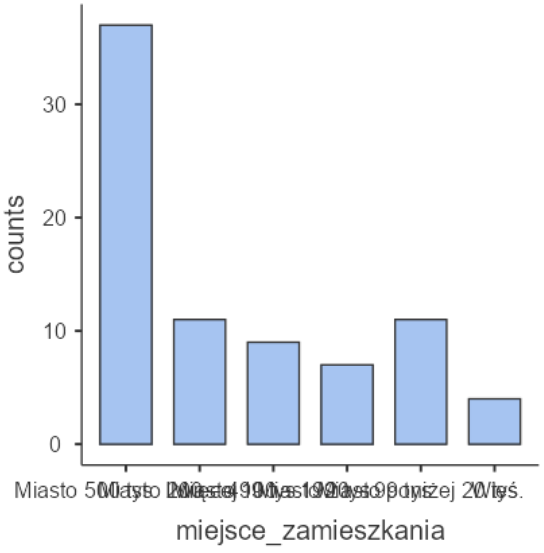


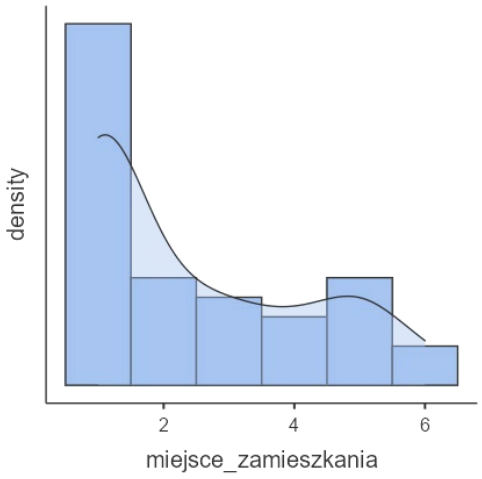


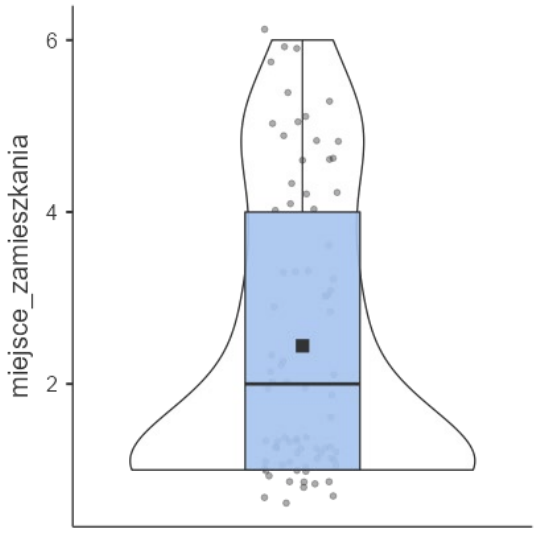


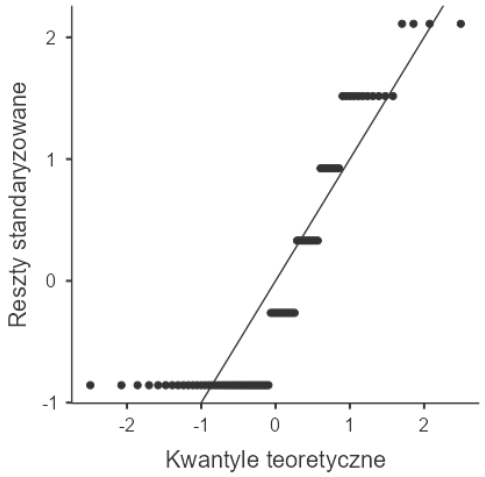


Deklarowane miejsce zamieszkania:

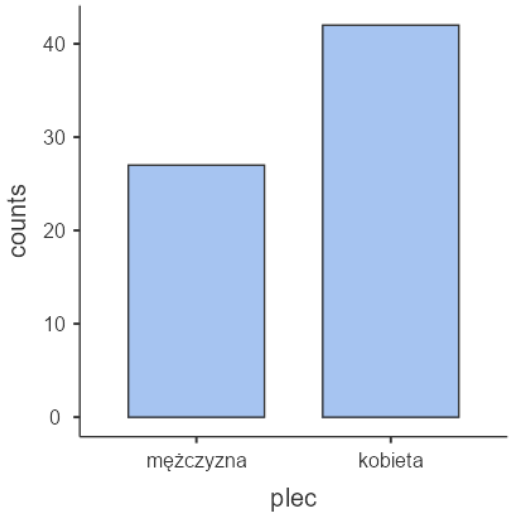


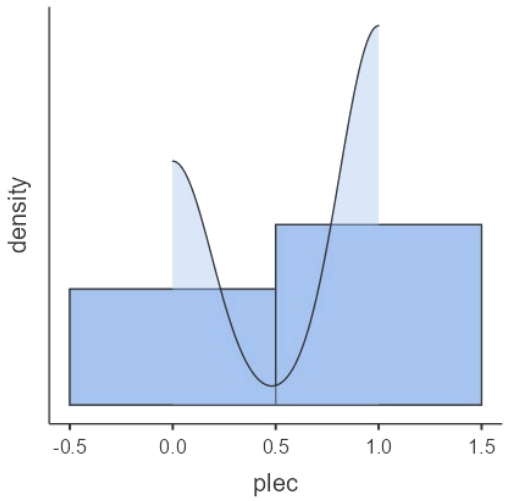


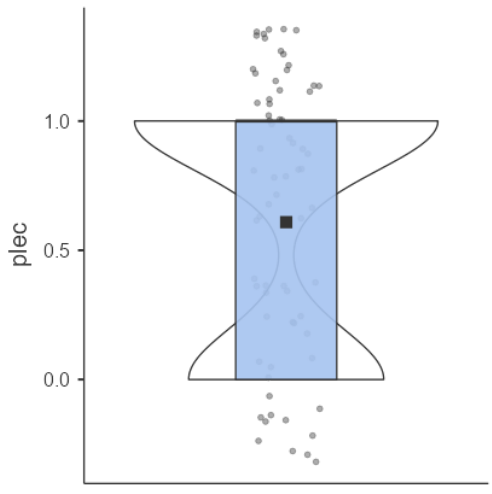


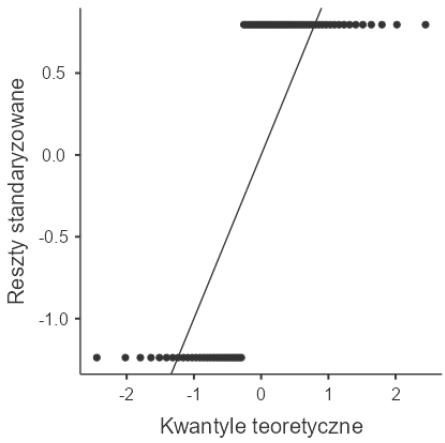


Deklarowana płeć:

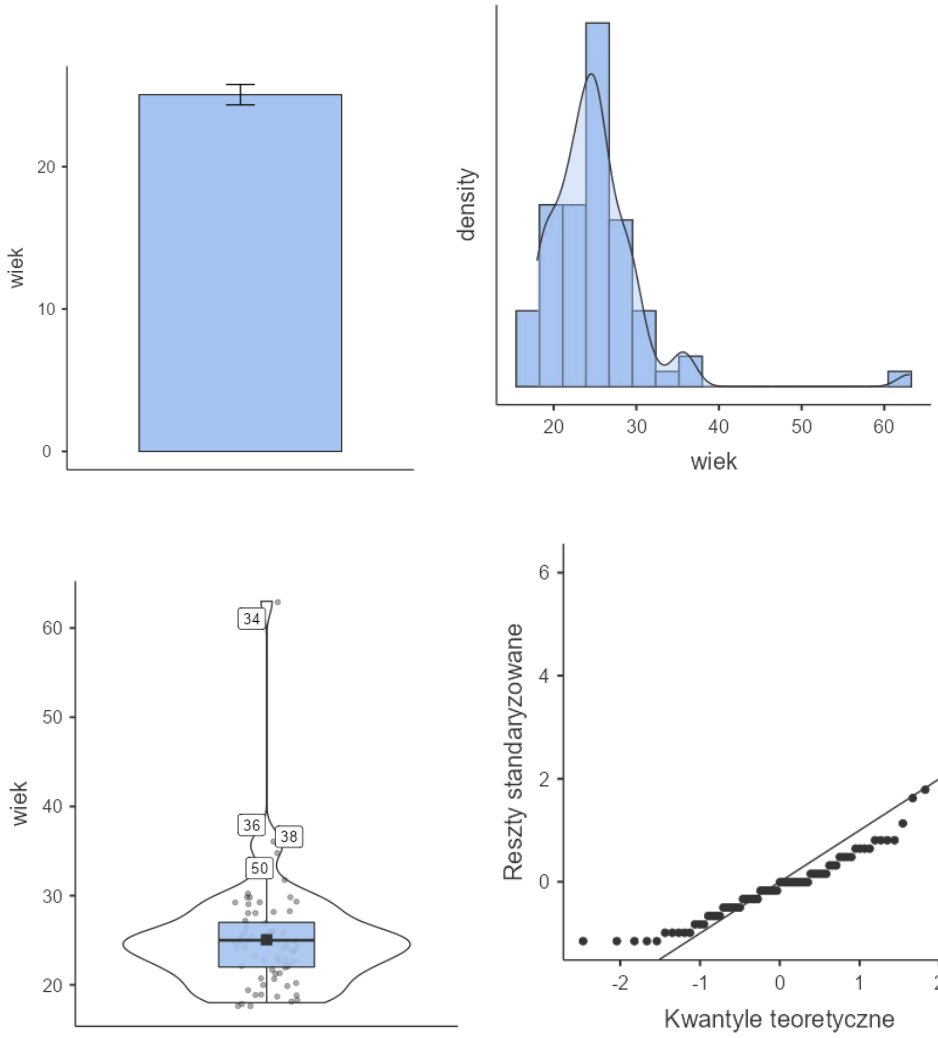




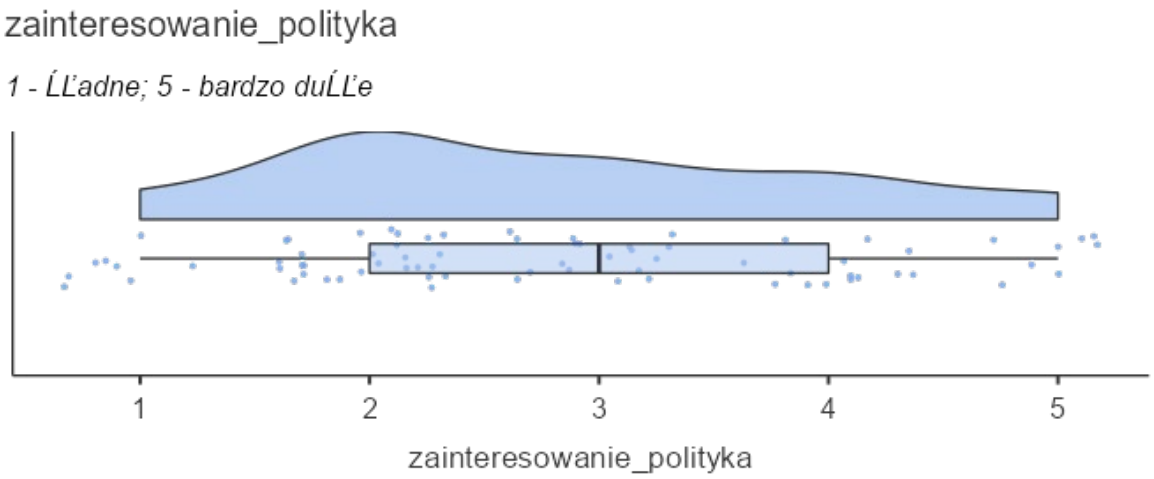




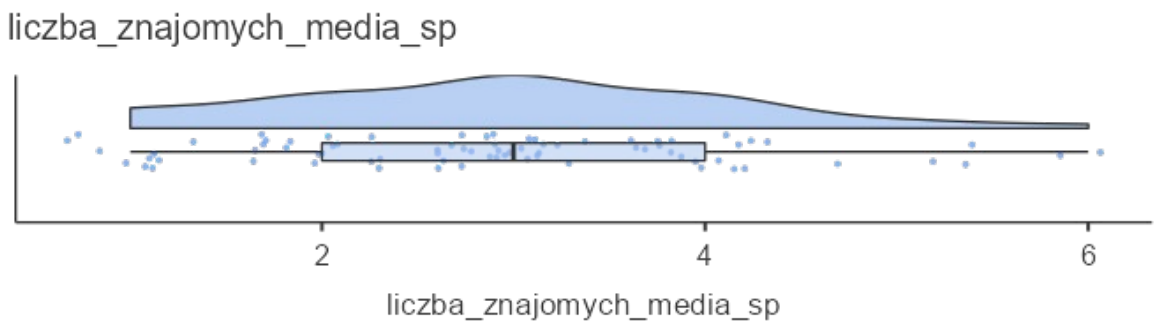
Deklarowany wiek:



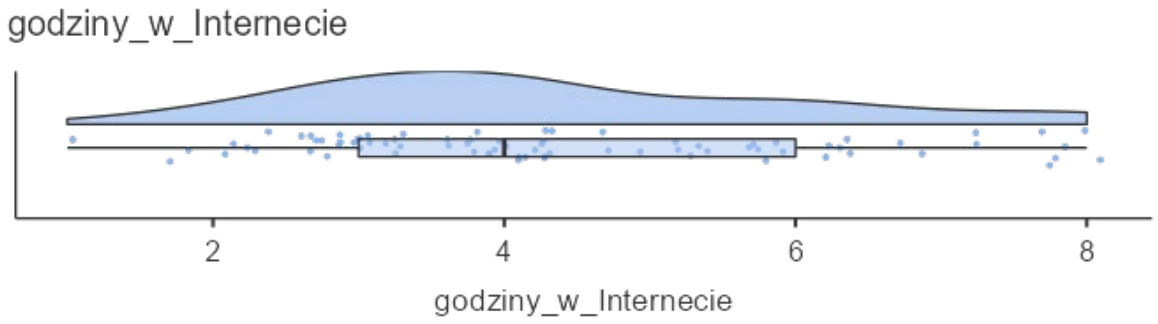
Deklarowane zainteresowanie polityką (1 – małe, 5 – duże):



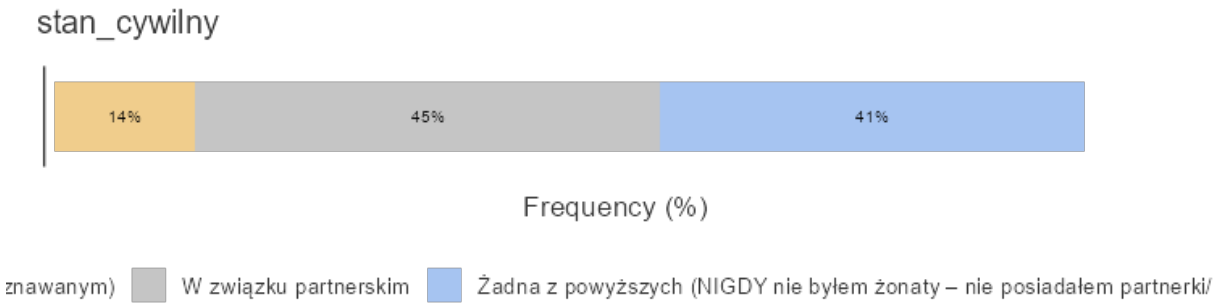
Liczba znajomych w mediach społecznościowych (1 – <100, 2 – 100-200, 3 – 200-300, 4 – 300-400, 5 – 400-500, 6 – 600<):



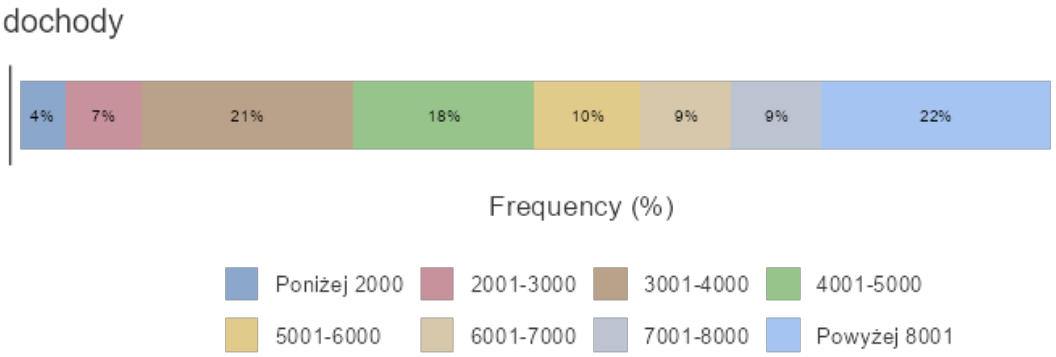
Deklarowana liczba godzin spędzana w Internecie:



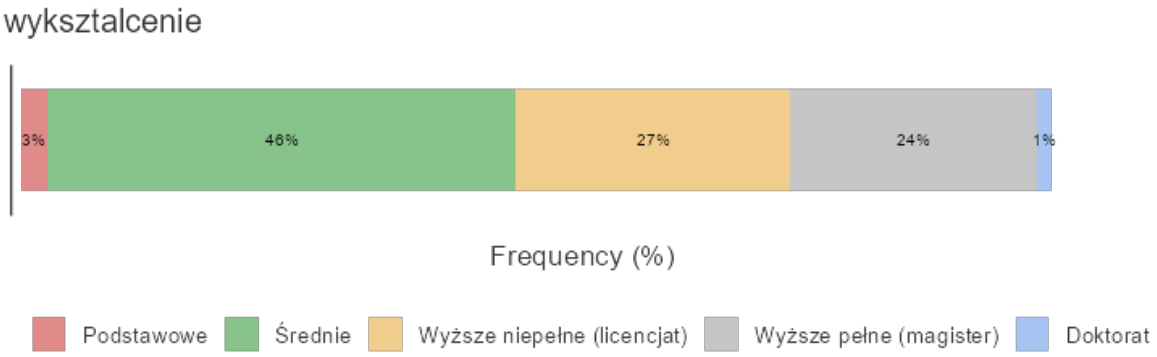
Deklarowany stan cywilny:



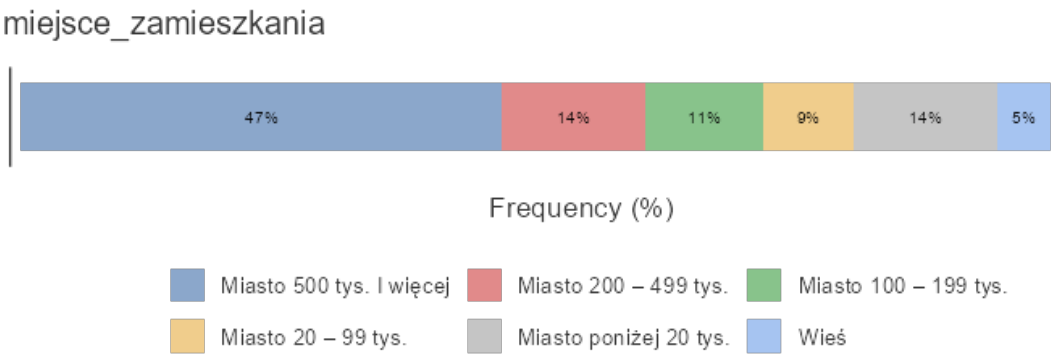
Deklarowane dochody:



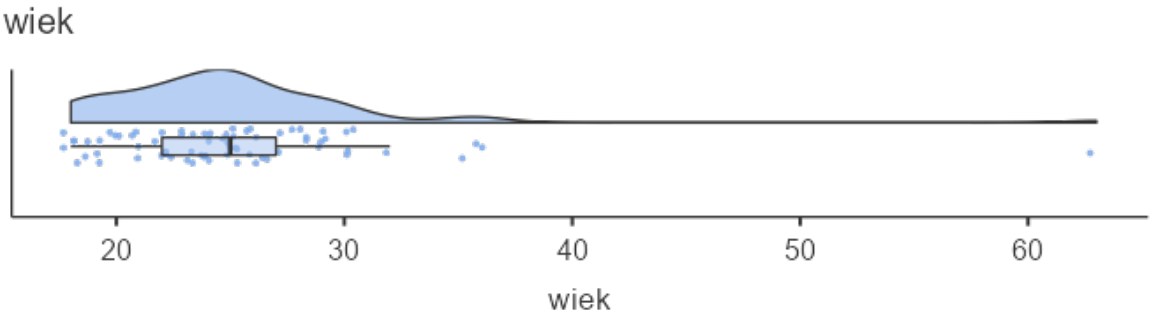
Deklarowane wykształcenie:



Deklarowane miejsce zamieszkania:

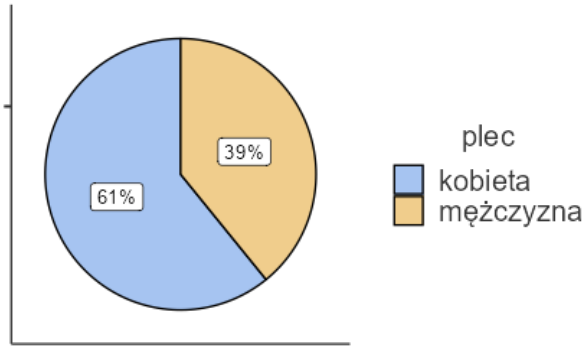


Deklarowany wiek:



Deklarowana płeć:

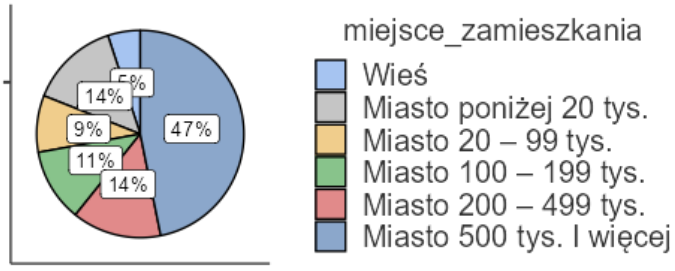
$\chi^2_{gof}(1) = 3.26, p = 0.071, \widehat{V}_{Cramer} = 0.22, CI_{95\%}$ [-0.(



favor of null: $\log_e(BF_{01}) = 0.71, a = 1.00$

Deklarowane miejsce zamieszkania:

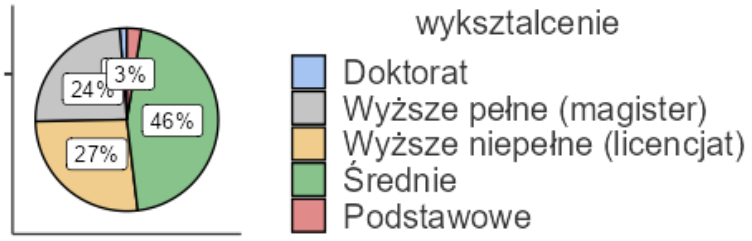
$\chi^2_{gof}(5) = 54.44, p = < 0.001, \widehat{V}_{Cramer} = 0.37, CI_{95\%}$ [0.2



: $\log_e(BF_{01}) = -9.45, a = 1.00$

Deklarowane wyksztalcenie:

$\chi^2_{gof}(4) = 54.10, p = < 0.001, \widehat{V}_{Cramer} = 0.41, CI_{95\%}$ [0.3



$BF_{01}) = -23.00, a = 1.00$

Załącznik 3 – Arkusz badawczy

Treść zgody na udział w badaniu:

Serdecznie dziękuję za zainteresowanie badaniem.

Niniejsze badanie ma na celu zbadanie najskuteczniejszej formy reklamy marketingowej – nagrań wideo – dla usług finansowych oraz oszacowania wpływu popularności wizerunku osób na odbiór kampanii.

W badaniu tym poproszę Cię o wypełnienie metryczki oraz pięciu krótkich kwestionariuszy, a następnie obejrzenie 6 różnych filmów, mających zachęcić Cię do inwestowania. Po każdym z filmów wyświetlone zostanie 17 krótkich pytań związanych z niniejszym badaniem.

Badanie jest w pełni anonimowe. Nie gromadzę żadnych danych pozwalających na identyfikację uczestników badania. Odpowiedzi, których udzielisz będą analizowane zbiorczo (razem z odpowiedziami innych osób), a nie indywidualnie i posłużą dla celów naukowych. Wzięcie udziału w badaniu oznacza wyrażenie zgody na badanie.

Badanie przeznaczone jest dla osób pełnoletnich.

Udział w badaniu jest w pełni dobrowolny. Możesz wycofać się z niego w dowolnym momencie – bez podawania jakiejkolwiek przyczyny.

Badanie prowadzone jest przez doktoranta Wojskowej Akademii Technicznej – Bartosza Bidermana. W razie pytań lub jakichkolwiek wątpliwości, proszę o wiadomość na adres: bartosz.biderman@wat.edu.pl.

6 losowo prezentowanych moderatorów:

moderator lęku – GAD 7 – 7 pytań,

moderator na depresyjność – PHQ-9 – 9 pytań,

moderator potrzeby poznawczego domknięcia – 15 pytań,

moderator kodów moralnych – MFQ – 30 pytań,

moderator samooceny – SES – 10 pytań,

moderator impulsywności – BIS-Brief – 8 pytań.

Losowe wyświetlanie 6 filmów, po każdym z nich jednakowe pytania w jednakowej kolejności, skala 1 – 10:

1	2	3	4	5	6	7	8	9	10

Proszę odpowiedz na poniższe pytania. (1 wcale, 10 bardzo dużo)

1. Na ile wyświetlony film wyglądał dla Ciebie naturalnie?
2. W jakim stopniu osoba, której wizerunek był prezentowany wzbudza twoje zaufanie?
3. Na ile jej filmowy przekaz wzbudza twoje zaufanie?
4. W jakim stopniu prezentowane nagranie zachęca Cię do inwestycji?
5. Jaki procent swoich oszczędności byłbyś skłonny zainwestować na polecanej platformie po obejrzeniu tego nagrania?
6. W jakim stopniu wierzysz w realność obiecywanego zysku?
7. Po obejrzeniu tego nagrania, w jakim stopniu obawiasz się utraty zainwestowanych środków na tej platformie?
8. W jakim stopniu film ten może wpłynąć w Twojej ocenie na decyzje inwestycje innych osób oglądających go?
9. W jakim stopniu uważasz, że ktoś może być skłonny do zainwestowania swoich środków po obejrzeniu tego nagrania?
10. Na ile przekonuje Cię prawdziwość powyższego nagrania?
11. Jak dobrze kojarzysz osobę prezentowaną na nagraniu?
12. Jak oceniasz wpływ powyższego nagrania na odbiór tej osoby przez znajomych / osoby obserwujące ją?
13. Czy znasz osobę wyświetlaną na nagraniu?
14. Czy polubił/a byś ten film?
15. Czy udostępnił/a byś ten film swoim znajomym?
16. Jakie elementy aktora / aktorki były dla ciebie nienaturalne (sztuczne) na powyższym nagraniu? [nieobowiązkowe, opisowe]
17. Co sądzisz o tym nagraniu? [nieobowiązkowe, opisowe]

Metryczka:

Twój wiek: [wpisanie liczby od 18 do 99]

Twoja płeć:

Mężczyzna

Kobieta

Twoje miejsce zamieszkania:

Miasto 500 tys. i więcej

Miasto 200 – 4999 tys.

Miasto 100 – 199 tys.

Miasto 20 – 99 tys.

Miasto poniżej 20 tys.

Wieś

Twoje wykształcenie:

Podstawowe

Średnie

Wyższe niepełne (licencjat)

Wyższe pełne (magister)

Doktorat

[jeżeli b lub c lub d lub e]

Proszę nazwij kierunek studiów na jaki uczęszczasz/ jaki ukończyłeś: [tekst]

Twoja sytuacja zawodowa:

a. Wykonuję pracę odpłatną (praca najemna, na własny rachunek, w firmie rodzinnej, we własnym gospodarstwie rolnym) – także w przypadku tymczasowej przerwy w pracy

b. Uczę się w szkole lub na uczelni (nauka nie jest opłacana przez pracodawcę) – także w przypadku wakacyjnej przerwy w nauce

c. Jestem bezrobotny/-a i aktywnie poszukuję pracy

d. Jestem bezrobotny/-a, nie poszukuję aktywnie pracy, ale chciałbym/-abym pracować

e. Trwała choroba lub niepełnosprawność

f. Jestem na emeryturze, rencie

g. Zajmuję się domem, opiekuję się dziećmi lub innymi osobami

h. (Inna sytuacja)

Które z poniższych określeń najlepiej opisuje rodzaj pracy wykonywanej przez Twojego ojca, kiedy miałeś/miałaś 14 lat?

a. Wolne zawody i specjaliści np.: lekarz – nauczyciel – inżynier – artysta – główny księgowy

b. Wyższe stanowiska administracyjne np.: bankowiec – dyrektor w dużej firmie – wysoki urzędnik państwowy – wyższy rangą działacz związkowy

c. Zawody związane z pracą biurową np.: sekretarz prezesa/dyrektora – urzędnik, referent – kierownik biura – pracownik działu księgowości

d. Sprzedaż, handel np.: kierownik działu sprzedaży – właściciel sklepu – sprzedawca – agent ubezpieczeniowy

e. Usługi np.: właściciel restauracji – policjant – kelner – fryzjer – żołnierz zawodowy

f. Wykwalifikowany pracownik fizyczny np.: brygadzista – mechanik samochodowy – drukarz – ślusarz narzędziowy – elektryk

g. Półwykwalifikowany pracownik fizyczny np.: murarz – kierowca autobusu – pracownik w przetwórstwie spożywczym– stolarz – blacharz – piekarz

h. Niewykwalifikowany pracownik fizyczny np.: robotnik – niewykwalifikowany pracownik w fabryce, magazynie – portier

i. Rolnictwo np.: rolnik – robotnik rolny – traktorzysta – rybak

j. Był bezrobotny

Które z poniższych określeń najlepiej opisuje rodzaj pracy wykonywanej przez Twoją matkę, kiedy miałeś/miałaś 14 lat?

a. Wolne zawody i specjaliści np.: lekarz – nauczyciel – inżynier – artysta – główny księgowy

b. Wyższe stanowiska administracyjne np.: bankowiec – dyrektor w dużej firmie – wysoki urzędnik państwowy – wyższy rangą działacz związkowy

c. Zawody związane z pracą biurową np.: sekretarz prezesa/dyrektora – urzędnik, referent – kierownik biura – pracownik działu księgowości

d. Sprzedaż, handel np.: kierownik działu sprzedaży – właściciel sklepu – sprzedawca – agent ubezpieczeniowy

e. Usługi np.: właściciel restauracji – policjant – kelner – fryzjer – żołnierz zawodowy

f. Wykwalifikowany pracownik fizyczny np.: brygadzista – mechanik samochodowy – drukarz – ślusarz narzędziowy – elektryk

g. Półwykwalifikowany pracownik fizyczny np.: murarz – kierowca autobusu – pracownik w przetwórstwie spożywczym– stolarz – blacharz – piekarz

h. Niewykwalifikowany pracownik fizyczny np.: robotnik – niewykwalifikowany pracownik w fabryce, magazynie – portier

i. Rolnictwo np.: rolnik – robotnik rolny – traktorzysta – rybak

j. Była bezrobotna

Jeśli dodasz dochody członków swojego gospodarstwa domowego ze wszystkich źródeł, która litera odpowiada dochodowi netto Twojego gospodarstwa na osobę? Jeśli nie znasz dokładnej ich wysokości, proszę je oszacować.

a. Poniżej 2000

b. 2001-3000

c. 3001-4000

d. 4001-5000

e. 5001-6000

f. 6001-7000

g. 7001-8000

h. Powyżej 8001

Które z poniższych określeń – tylko jedno – opisuje Twój obecny stan cywilny?

a. Pozostający w związku małżeńskim (prawnie uznawanym)

b. W związku partnerskim

c. Rozwiedziony/-a (rozwód orzeczony przez sąd)

d. Wdowiec/wdowa

e. Żadna z powyższych (NIGDY nie byłem żonaty – nie posiadałem partnerki/ NIGDY nie byłam zamężna – nie posiadałam partnera)

Oszacuj proszę średnią ilość godzin spędzaną dziennie w Internecie (zawierając w tym czas spędzony zarówno na komputerze jak i komórce)

a. Poniżej 1h

b. 1-2h

c. 3-4h

d. 5-6h

e. 7-8h

f. 8-9h

g. 9-10h

h. Powyżej 10h

Na którą partię lub ugrupowanie głosowałeś/głosowałaś w ostatnich wyborach do Sejmu?

a. KOMITET WYBORCZY POLSKIE STRONNICTWO LUDOWE

b. KOMITET WYBORCZY PRAWO i SPRAWIEDLIWOŚĆ

c. KOMITET WYBORCZY SOJUSZ LEWICY DEMOKRATYCZNEJ

d. KOMITET WYBORCZY KONFEDERACJA WOLNOŚĆ i NIEPODLEGŁOŚĆ

e. KOALICYJNY KOMITET WYBORCZY KOALICJA OBYWATELSKA PO. N IPL ZIELONI

f. KOMITET WYBORCZY WYBORCÓW KOALICJA BEZPARTYJNI i SAMORZĄDOWCY

g. Nie głosowałem/łam

Na jaki komitet zagłosowałbyś/zagłosowałabyś, gdyby wybory odbyły się za miesiąc?

a. KOMITET WYBORCZY POLSKIE STRONNICTWO LUDOWE

b. KOMITET WYBORCZY PRAWO i SPRAWIEDLIWOŚĆ

c. KOMITET WYBORCZY SOJUSZ LEWICY DEMOKRATYCZNEJ

d. KOMITET WYBORCZY KONFEDERACJA WOLNOŚĆ i NIEPODLEGŁOŚĆ

e. KOALICYJNY KOMITET WYBORCZY KOALICJA OBYWATELSKA PO. N IPL ZIELONI

f. KOMITET WYBORCZY WYBORCÓW KOALICJA BEZPARTYJNI i SAMORZĄDOWCY

g. Nie głosowałbym/ głosowałabym

Oszacuj proszę przybliżoną liczbę znajomych, których masz w swojej sieci w mediach społecznościowych, z których korzystasz najczęściej.

(1 = 0-50 osób, 2 = 51-200 osób, 3 = 201-500 osób, 4 = 501–1000 osób, 5 = 1001–2000 osób i 6 = ponad 2000 osób)

Określ proszę na poniższej skali swoje zainteresowanie polityką.

(1 = żadne, 5 = bardzo duże)

Treść odkłamania:

Dziękuję Ci serdecznie za wzięcie udziału w badaniu.

Rzeczywistym celem badania było sprawdzenie, czy rozpoznasz fałszywe nagrania, zmanipulowane, w których prezentowani celebryci w rzeczywistości nie brali udziału. Powstały one przy wykorzystaniu technologii deepfake, która pozwala na podszywanie się pod wizerunek innych osób.

Deepfake to (zbitka wyrazowa od ang. deep learning „głębokie uczenie" oraz fake „fałszywy") – technika obróbki obrazu, polegająca na łączeniu obrazów twarzy ludzkich przy użyciu technik sztucznej inteligencji.

Technika stosowana jest do łączenia i nakładania obrazów nieruchomych i ruchomych na obrazy lub filmy źródłowe przy użyciu komputerowych systemów uczących się. Uzyskane w ten sposób łudząco realistyczne ruchome obrazy stosowane są

w nagraniach filmowych, stwarzając możliwości manipulacji poprzez np. niemożliwą do odróżnienia przez widza zamianę twarzy aktorów występujących w filmie. Może to prowadzić do skompromitowania osoby poprzez fałszywe „obsadzenie jej" np. W roli w filmie pornograficznym. Jednym z zagrożeń fałszywymi informacjami jest możliwość wpłynięcia na wyniki wyborów.

Przepraszam, że wprowadziłem Cię w błąd co do celu badania, jednak jest to często stosowana metoda, dzięki której w eksperymencie można określić przyczyny zjawisk. Nie musisz się martwić, jeżeli nieznana Ci była ta technologia albo jeżeli nie rozpoznałeś fałszywych nagrań. Dużo osób wciąż nie zdaje sobie sprawy z rosnącego zagrożenia filmami deepfake.

Proszę nie mów swoim znajomym o faktycznym celu badania.

Jeżeli interesują Cię wyniki badania, proszę napisz do mnie maila – bartosz.biderman@wat.edu.pl

Ankieta końcowa:

a) Czy byłeś/aś świadomy/a istnienia nagrań deepfake przed tą ankietą?

b) Czy kiedykolwiek przypadkowo udostępniłeś/aś film deepfake, który później okazał się być oszustwem? (0 = nigdy, 1 = raz, 2 = więcej niż raz)

c) Czy zdarzyło Ci się natrafić na filmy deepfake w mediach społecznościowych? (1 = wcale, 5 = bardzo dużo)

d) Czy zdarzało Ci się natrafić na filmy deepfake ogólnie w Internecie? (1 = wcale, 5 = bardzo dużo)

e) Jak uważasz, czy w niniejszym badaniu udało Ci się rozpoznać nagrania przygotowane technologią deepfake? (tak, nie, nie wiem)

f) Czy uważasz, że jesteś w stanie rozpoznawać, iż dane nagranie znalezione w Internecie jest fałszywe i zostało przygotowane przy wykorzystaniu deepfake? (tak, nie, nie wiem)

g) w jakim stopniu jesteś zaniepokojony/a dalszym rozwojem technologii deepfake? (1 = wcale do 5 = bardzo)

Załącznik 4 – Skróty statystyczne użyte w pracy

- BIS-Brief – moderator impulsywności
- *D* – dominanta
- GAD – moderator lęku
- *K* – kurtoza (standaryzowany moment średniej czwartego rzędu)
- *M* – wartość średniej
- *Max* – maksymalna wartość odpowiedzi
- *Me* – mediana
- MFQ – moderator kodów moralnych
- *Min* – minimalna wartość odpowiedzi
- N – liczba osób badanych, które odpowiedziały na dane pytanie
- PDP – moderator potrzeby domknięcia poznawczego
- PHQ – moderator na depresyjność
- $p_{S\text{-}W}$ – istotność statystyczna testu Shapiro-Wilk
- rho (ρ) – Rho Spearmana
- *SD* – odchylenie standardowe
- *SE* – błąd standardowy średniej
- SE_K (Std. error K) – błąd standardowy kurtozy
- SES – moderator samooceny
- SE_{SKE} – błąd standardowy skośności
- *SKE* – skośność (trzeci moment centralny)
- *S-W* – wartość testu Shapiro-Wilk