ANALIZA TREŚCI AUDIOWIZUALNYCH JAKO ELEMENT SYSTEMÓW BEZPIECZEŃSTWA PAŃSTWA



Marta Chodyka
Zbigniew Ciekanowski

Akademia Bialska im. Jana Pawła II

ANALIZA TREŚCI AUDIOWIZUALNYCH JAKO ELEMENT SYSTEMÓW BEZPIECZEŃSTWA PAŃSTWA

Marta Chodyka

Zbigniew Ciekanowski

Biała Podlaska 2024

Wydawca
Akademia Bialska im. Jana Pawła II

Recenzja
prof. dr hab. inż. Marian Kopczewski, Akademia Wojsk Lądowych
dr hab. Julia Nowicka, Akademia Sztuki Wojennej

© Copyright by Akademia Bialska im. Jana Pawła II

Książka jest udostępniana na licencji Creative Commons Uznanie autorstwa – Użycie niekomercyjne – Bez utworów zależnych 4.0 Międzynarodowa

ISBN 978-83-68103-18-2

Wydawnictwo AB JPII
ul. Sidorska 95/97
21-500 Biała Podlaska
wydawnictwo.akademiabialska.pl

Projekt okładki, skład

druk-24h.com.pl
DRUKARNIA CYFROWA
Hurtowa 21, 15-399 Białystok
tel. 535 500 898
e-mail: studio@druk-24h.com.pl

Spis treści

Wstęp 6

1. Wybrane zagadnienia związane z funkcjonowaniem systemu bezpieczeństwa państwa 13

1.1. System bezpieczeństwa narodowego 13

1.2. Struktura i główne podmioty systemu bezpieczeństwa narodowego 18

1.3. Mechanizmy zarządzania bezpieczeństwem państwa - strategiczne dokumenty i polityki 23

1.4. Informacja i komunikacja jako element bezpieczeństwa 26

2. System bezpieczeństwa informacyjnego państwa 32

2.1. Pojęcie i zakres bezpieczeństwa informacyjnego 32

2.2. Struktura systemu bezpieczeństwa informacyjnego 37

2.3. Kluczowe podmioty systemu bezpieczeństwa informacyjnego 39

2.4. Zagrożenia dla bezpieczeństwa informacyjnego 47

3. Rola analizy treści audiowizualnych w systemach bezpieczeństwa państwa 51

3.1. Znaczenie analizy treści audiowizualnych 51

3.2. Multimodalność w analizie zagrożeń informacyjnych 55

3.3. Wpływ technologii cyfrowych na bezpieczeństwo informacyjne państwa 59

3.4. Koncepcja wojny informacyjnej w kontekście audiowizualnym 65

4. Współczesne zagrożenia dezinformacyjne i medialne 71
4.1. Rola platform cyfrowych i algorytmów w rozpowszechnianiu dezinformacji 71
4.2. *Fake newsy* i mechanizmy manipulacji audiowizualnej 73
4.3. Deepfake jako narzędzie destabilizacji politycznej i militarnej 76
4.4. Analiza przypadków wpływu dezinformacji audiowizualnej na bezpieczeństwo państwa 80
4.5. Strategie przeciwdziałania dezinformacji w różnych państwach 91

5. Technologie analizy treści audiowizualnych 100
5.1. Metody wykrywania zmanipulowanych treści audiowizualnych .. 100
5.2. Algorytmy analizy obrazu i dźwięku w detekcji dezinformacji 104
5.3. Systemy monitorowania wideo w czasie rzeczywistym 108
5.4. Detekcja *deepfake* – skuteczność i ograniczenia 114
5.5. Uwarunkowania prawne i stanowiska międzynarodowe w kwestii monitorowania treści 124

6. Zastosowanie algorytmów analizy treści audiowizualnych w bezpieczeństwie narodowym 128
6.1. Algorytmy rozpoznawania obiektów audiowizualnych 128
6.2. Porównanie klasycznych i nowoczesnych metod analizy wideo .. 134
6.3. Integracja systemów wykrywania dezinformacji z instytucjami bezpieczeństwa 141
6.4. Przykłady praktycznego wykorzystania narzędzi analizy audiowizualnej w bezpieczeństwie narodowym 155

7. Badania empiryczne – skuteczność metod analizy treści audiowizualnych 173
7.1. Opis zastosowanej metodologii badawczej 173
7.2. Eksperymenty na zbiorach danych audiowizualnych – testowanie skuteczności algorytmów 175
7.3. Analiza wyników badań – porównanie wykrywalności manipulacji 178
7.4. Wpływ zastosowanych metod na zwiększenie odporności państwa na dezinformację 181

8. Wyzwania i przyszłość technologii wykrywania dezinformacji 189
8.1. Ograniczenia technologiczne w analizie treści audiowizualnych 189
8.2. Etyczne aspekty automatycznego wykrywania manipulacji medialnych 195
8.3. Rola regulacji prawnych w rozwoju technologii analizy 200
8.4. Rozwój AI w analizie treści audiowizualnych i jego konsekwencji 204
8.5. Interdyscyplinarność w rozwoju narzędzi do wykrywania dezinformacji 209

9. Wnioski, rekomendacje i kierunki dalszych działań i badań 212
9.1. Rekomendacje dla polityki bezpieczeństwa państwa 212
9.2. Rekomendacje dla instytucji bezpieczeństwa i mediów publicznych 218
9.3. Edukacja medialna i cyfrowa jako narzędzie wzmacniania odporności społecznej 220
9.4. Możliwości dalszych badań nad dezinformacją audiowizualną 223
9.5. Strategie walki z *deepfake* i manipulacją medialną na poziomie państwowym 226
9.6. Wnioski 230

10. Podsumowanie 233

11. Literatura 235

12. Wykaz rysunków 248

13. Wykaz tabel 249

Wstęp

Współczesne państwa funkcjonują w warunkach dynamicznego rozwoju technologii cyfrowych, które wpływają na szereg istotnych zmian w obszarze bezpieczeństwa narodowego. Jeszcze do niedawna bezpieczeństwo utożsamiano głównie z kwestiami militarnymi, gospodarczymi lub politycznymi, jednak współcześnie coraz bardziej dostrzegalna staje się rola sfery informacyjnej, która obejmuje nowe rodzaje zagrożeń wynikających z postępującej cyfryzacji społeczeństwa. Upowszechnienie się Internetu, gwałtowny wzrost popularności platform społecznościowych oraz wszechobecność mediów elektronicznych sprawiają, że państwa muszą mierzyć się z zagrożeniami związanymi z rozpowszechnianiem fałszywych treści informacyjnych oraz manipulacją opinią publiczną. W szczególności, w obszarze bezpieczeństwa informacyjnego dużą rolę zaczęły odgrywać treści audiowizualne, które są obecnie głównym nośnikiem komunikacji społecznej i politycznej. Dzięki zaawansowanym technologiom cyfrowym, materiały audiowizualne – obrazy, nagrania audio oraz materiały wideo – docierają do milionów odbiorców na całym świecie w bardzo krótkim czasie, wywierając przy tym istotny wpływ na postrzeganie rzeczywistości przez społeczeństwo, kształtowanie nastrojów społecznych oraz procesy decyzyjne o charakterze politycznym.

Rosnąca popularność materiałów audiowizualnych w komunikacji masowej wynika przede wszystkim z ich wyjątkowej siły oddziaływania na ludzką percepcję. Psychologiczne mechanizmy, na których bazują przekazy wizualne i dźwiękowe, opierają się na naturalnej podatności ludzkiego umysłu na bodźce audiowizualne, które są odbierane szybciej, intuicyjnie oraz mniej krytycznie niż przekazy tekstowe. Liczne badania wskazują, że obraz oraz dźwięk skuteczniej angażują odbiorców emocjonalnie, wywołując natychmiastową reakcję psychologiczną oraz obniżając skłonność do refleksji krytycznej. W rezultacie pojawia się ryzyko celowej manipulacji

treściami audiowizualnymi, które mogą być wykorzystywane zarówno do osiągania celów komercyjnych, jak i politycznych lub nawet militarnych. Manipulacje te przybierają różne formy – od stosunkowo prostych montażowych zmian kontekstu obrazu czy dźwięku, aż po zaawansowane techniki wykorzystujące sztuczną inteligencję, takie jak tzw. deepfake, które pozwalają na generowanie realistycznych, lecz całkowicie fikcyjnych materiałów audiowizualnych. Ze względu na wysoki poziom realizmu, takie manipulacje audiowizualne mogą wprowadzać w błąd szeroką grupę odbiorców, powodując chaos informacyjny, niepokoje społeczne lub destabilizację polityczną. W obliczu narastających zagrożeń informacyjnych związanych z treściami audiowizualnymi, kluczowe znaczenie dla bezpieczeństwa państwa zyskuje rozwój skutecznych metod i narzędzi analitycznych, pozwalających na systematyczne wykrywanie i neutralizację tego rodzaju manipulacji. Niniejsza monografia podejmuje właśnie tę problematykę, stawiając sobie za cel kompleksową analizę współczesnych metod identyfikacji zmanipulowanych materiałów audiowizualnych, ocenę ich skuteczności oraz możliwości praktycznego wykorzystania w działaniach instytucji odpowiedzialnych za bezpieczeństwo informacyjne państwa. U podstaw przyjętych założeń leży przekonanie, że skuteczne przeciwdziałanie dezinformacji audiowizualnej musi być podejściem interdyscyplinarnym, uwzględniającym nie tylko aspekty technologiczne, ale także społeczne, prawne oraz psychologiczne.

Analiza treści audiowizualnych, stanowiąca przedmiot niniejszej monografii, to złożony proces badawczy, którego celem jest systematyczne i możliwie obiektywne ocenianie materiałów w formie zdjęć, filmów oraz nagrań dźwiękowych pod kątem ich autentyczności, spójności treściowej i intencji ich twórców. Specyfika analizy audiowizualnej wynika z faktu, że współczesne przekazy medialne mają charakter wielomodalny – zawierają jednocześnie obraz, dźwięk, a nierzadko także tekst – co wymaga stosowania zaawansowanych narzędzi analitycznych, często opartych na najnowszych osiągnięciach informatyki, w szczególności sztucznej inteligencji oraz uczenia maszynowego. Ze względu na swoją wielowymiarowość, analiza taka nie może opierać się jedynie na klasycznych metodach oceny treści tekstowych, lecz musi integrować podejścia właściwe naukom technicznym, społecznym oraz humanistycznym. Oznacza to konieczność stosowania zarówno zaawansowanych technik komputerowego rozpoznawania obrazów, algorytmów detekcji manipulacji dźwiękowej, jak i metod interpretacyjnych

uwzględniających kontekst społeczny, polityczny oraz kulturowy analizowanych materiałów. Treści audiowizualne stanowią przy tym coraz częściej element skoordynowanych kampanii manipulacyjnych, mających na celu destabilizację wewnętrzną państw lub podważanie zaufania społecznego do kluczowych instytucji. Manipulacje tego rodzaju są szczególnie niebezpieczne, ponieważ za pomocą realistycznie przygotowanych materiałów audiowizualnych można szybko i skutecznie rozpowszechniać fałszywe narracje, które błyskawicznie zdobywają popularność w mediach społecznościowych. Szczególnie groźnym i jednocześnie istotnym dla analizowanych zagadnień zjawiskiem są materiały typu *deepfake*, które generowane są z wykorzystaniem zaawansowanych algorytmów uczenia głębokiego – głównie w oparciu o generatywne sieci przeciwstawne (GAN). Metody te pozwalają na syntetyczne tworzenie nagrań przedstawiających realistyczne, lecz całkowicie fikcyjne wypowiedzi osób publicznych, co może prowadzić do poważnych kryzysów informacyjnych, społecznych i politycznych. W efekcie odbiorcy, mając ograniczone możliwości weryfikacji autentyczności takich nagrań, często reagują w sposób emocjonalny, przyjmując zmanipulowane treści jako prawdziwe, co z kolei sprzyja dalszemu rozpowszechnianiu tych fałszywych materiałów. Badania nad *deepfake* wskazują jednoznacznie, że rosnący poziom realizmu uzyskiwanego przez algorytmy generatywne powoduje, iż coraz trudniej jest odróżnić materiały autentyczne od zmanipulowanych. Z tego względu kluczowe jest rozwijanie i implementacja zaawansowanych technologicznie metod weryfikacji treści audiowizualnych, aby przeciwdziałać zjawisku dezinformacji oraz ograniczać jego negatywne skutki dla bezpieczeństwa państwa.

Z uwagi na przedstawione powyżej wyzwania, głównym celem badawczym niniejszej monografii jest dokonanie kompleksowej, pogłębionej analizy problematyki wykrywania manipulacji audiowizualnych, ze szczególnym uwzględnieniem oceny efektywności dostępnych obecnie technologii oraz metod analitycznych pod kątem ich praktycznego zastosowania w systemach bezpieczeństwa państwowego. Realizacja tego celu wymaga szczegółowego zbadania kilku istotnych zagadnień badawczych. Po pierwsze, konieczne jest zidentyfikowanie i precyzyjne scharakteryzowanie rodzajów zagrożeń wynikających ze świadomego i celowego manipulowania przekazami audiowizualnymi, z naciskiem na takie działania, które niosą szczególnie duże ryzyko destabilizacji społecznej, kryzysów politycznych

oraz realnego zagrożenia dla bezpieczeństwa wewnętrznego państwa. Po drugie, przeprowadzona analiza musi uwzględniać dogłębne porównanie dostępnych metod wykrywania manipulacji w przekazach audiowizualnych, z priorytetowym traktowaniem rozwiązań opartych na głębokich sieciach neuronowych, metodach uczenia maszynowego oraz podejściach multimodalnych, czyli takich, które jednocześnie integrują analizę obrazu, dźwięku oraz elementów tekstowych. Po trzecie, niezbędne jest określenie sposobów, w jakie najskuteczniejsze zidentyfikowane narzędzia i metody analityczne mogłyby być efektywnie włączone do struktur instytucji odpowiedzialnych za bezpieczeństwo państwa. Ważnym aspektem prowadzonych analiz jest także opracowanie szczegółowych rekomendacji dotyczących praktycznych warunków wdrożenia takich rozwiązań, uwzględniających zarówno aspekty technologiczne, jak i organizacyjne, prawne oraz społeczne. Tym samym nadrzędnym celem monografii jest stworzenie spójnego, interdyscyplinarnego modelu przeciwdziałania zagrożeniom informacyjnym wynikającym z manipulacji audiowizualnych, który może stanowić podstawę dla polityk publicznych oraz praktyki bezpieczeństwa informacyjnego.

Metodologia zastosowana w niniejszej monografii opiera się na interdyscyplinarnym podejściu, łączącym metody i narzędzia charakterystyczne dla nauk o bezpieczeństwie, medioznawstwa, informatyki, prawa oraz psychologii społecznej. W pierwszym etapie badań przeprowadzono obszerny przegląd literatury przedmiotu, obejmujący zarówno krajowe, jak i zagraniczne publikacje naukowe, koncentrujące się na teorii dezinformacji, technikach propagandy audiowizualnej, technologiach wykorzystywanych do manipulacji medialnych oraz strategiach zwalczania zagrożeń informacyjnych. Przegląd ten umożliwił określenie aktualnego stanu badań, identyfikację kluczowych problemów oraz wskazanie obszarów wymagających dalszego pogłębienia analitycznego. Uzupełnieniem analizy literaturowej była szczegółowa analiza dostępnych raportów i dokumentów strategicznych pochodzących zarówno od instytucji krajowych, jak i międzynarodowych, takich jak Unia Europejska, Organizacja Traktatu Północnoatlantyckiego (NATO), czy agencje specjalizujące się w monitoringu przestrzeni informacyjnej. Dokumenty te zostały poddane szczegółowej analizie jakościowej, co umożliwiło ocenę skuteczności dotychczas stosowanych praktyk w zakresie ochrony przestrzeni informacyjnej oraz identyfikację dobrych praktyk i możliwych rekomendacji wdrożeniowych. W drugim, empirycznym etapie

prowadzonych badań skoncentrowano się na praktycznym porównaniu efektywności wybranych algorytmów i metod wykorzystywanych do automatycznego wykrywania manipulacji treści audiowizualnych, w szczególności materiałów określanych mianem *deepfake*. Przygotowano specjalne zestawy materiałów testowych, które obejmowały syntetycznie wygenerowane nagrania audio oraz wideo, w sposób kontrolowany i systematyczny zmodyfikowane przy użyciu technik manipulacji cyfrowej. Materiały te zawierały realistyczne przykłady fałszerstw wizualnych oraz dźwiękowych, symulujące rzeczywiste sytuacje, z którymi mogą się mierzyć służby odpowiedzialne za bezpieczeństwo informacyjne. W eksperymentach wykorzystano szeroki wachlarz dostępnych metod, w tym algorytmy oparte na głębokich sieciach neuronowych, analizy multimodalne (jednoczesne analizowanie obrazu, dźwięku oraz tekstu), jak również metody tradycyjne, oparte na klasycznych analizach forensycznych. Eksperymentalna weryfikacja efektywności poszczególnych metod pozwoliła nie tylko na obiektywne porównanie ich skuteczności, lecz również umożliwiła identyfikację konkretnych ograniczeń technologicznych i metodologicznych, wskazując obszary wymagające dalszego rozwoju technologicznego. Równolegle z badaniami empirycznymi prowadzono badania dokumentacyjne, obejmujące szczegółową analizę regulacji prawnych oraz rozwiązań politycznych i organizacyjnych stosowanych w wybranych państwach europejskich, w tym we Francji, Niemczech oraz Wielkiej Brytanii. Przedmiotem analizy były szczególnie te regulacje, które wprowadzają ograniczenia związane z rozpowszechnianiem zmanipulowanych treści audiowizualnych, ze szczególnym naciskiem na ich stosowanie w kontekście wyborów, kampanii politycznych oraz kryzysów informacyjnych. Analiza ta miała na celu ocenę realnej skuteczności stosowanych rozwiązań legislacyjnych oraz instytucjonalnych, jak również zidentyfikowanie ewentualnych rekomendacji mogących stanowić wartościowy wkład w proces tworzenia podobnych mechanizmów ochronnych w polskim systemie bezpieczeństwa informacyjnego.

Wyniki przeprowadzonych badań wskazują na wysoki poziom skuteczności współczesnych metod analizy treści audiowizualnych, zwłaszcza tych, które wykorzystują zaawansowane techniki uczenia głębokiego oraz podejście integrujące różne modalności przekazu. Szczególną efektywnością cechują się rozwiązania bazujące na głębokich sieciach neuronowych, pozwalające na precyzyjne wykrywanie nawet subtelnych modyfikacji

i manipulacji materiałów audiowizualnych. Jednakże, mimo zaawansowania technologicznego stosowanych narzędzi, badania ukazały również istotne ograniczenia wynikające z ciągłego postępu w obszarze technologii generatywnych, które umożliwiają coraz bardziej wyrafinowane manipulacje audiowizualne, coraz trudniejsze do wykrycia za pomocą istniejących algorytmów analitycznych. Wnioski z badań wskazują na potrzebę ciągłego rozwijania technologii detekcyjnych, zwłaszcza poprzez integrację analizy multimodalnej oraz stosowanie hybrydowych modeli analitycznych, uwzględniających zarówno automatyczną analizę komputerową, jak i ekspercką interpretację kontekstową materiałów. Praktyczne wnioski zawarte w niniejszej monografii mogą znaleźć szerokie zastosowanie przede wszystkim w działaniach służb specjalnych, instytucji państwowych odpowiedzialnych za bezpieczeństwo wewnętrzne, sektora medialnego oraz organizacji pozarządowych zajmujących się monitoringiem mediów. Uzyskane wyniki mogą służyć jako istotna podstawa zarówno do wdrażania nowoczesnych systemów analizy treści audiowizualnych, jak również opracowywania strategii edukacyjnych i szkoleniowych, których celem jest wzmacnianie kompetencji cyfrowych obywateli oraz podnoszenie świadomości społecznej dotyczącej zagrożeń wynikających z manipulacji informacyjnej. Ponadto, wyniki monografii mogą być punktem wyjścia do kolejnych badań naukowych, w szczególności tych skupiających się na rozwoju jeszcze bardziej zaawansowanych technologii detekcji, doskonaleniu metod analitycznych oraz pogłębianiu wiedzy na temat psychologicznych mechanizmów podatności na dezinformację audiowizualną. Tym samym praca ta stanowi kompleksowe opracowanie problematyki analizy treści audiowizualnych w kontekście współczesnych wyzwań dla bezpieczeństwa informacyjnego państw.

Podsumowując, niniejsza monografia stanowi kompleksowe opracowanie problematyki analizy treści audiowizualnych, ukazując jej szczególną wagę dla bezpieczeństwa współczesnych państw w kontekście rosnących zagrożeń związanych z manipulacjami medialnymi i dezinformacją. Przeprowadzona analiza literatury naukowej pozwoliła na identyfikację kluczowych zagadnień oraz na umiejscowienie badań własnych na tle istniejących opracowań. Dzięki zastosowaniu metodologii interdyscyplinarnej oraz przeprowadzeniu szczegółowych badań empirycznych, obejmujących m.in. testowanie skuteczności zaawansowanych algorytmów opartych na głębokim uczeniu, udało się precyzyjnie określić mocne strony oraz ograniczenia

obecnie dostępnych technologii detekcyjnych. Wnioski wynikające z tych badań umożliwiły sformułowanie praktycznych rekomendacji dla instytucji państwowych, służb odpowiedzialnych za bezpieczeństwo wewnętrzne, sektora medialnego oraz organizacji zajmujących się edukacją medialną społeczeństwa. Tym samym praca wnosi istotny wkład nie tylko do nauk o bezpieczeństwie, ale także do medioznawstwa, informatyki oraz szeroko rozumianej praktyki bezpieczeństwa informacyjnego, wskazując jednocześnie na dalsze, obiecujące kierunki rozwoju badań naukowych w tej dziedzinie.

I

Wybrane zagadnienia związane z funkcjonowaniem systemu bezpieczeństwa państwa

1.1. System bezpieczeństwa narodowego

Priorytetowym interesem Rzeczypospolitej Polskiej pozostaje zapewnienie bezpieczeństwa narodowego. W obliczu zagrożeń cechujących się dużą zmiennością i skomplikowaną naturą konieczne jest stworzenie zintegrowanego systemu, którego podstawowym celem jest skuteczne przeciwdziałanie wszelkim potencjalnym niebezpieczeństwom[1]. Dlatego zasadnym wydaje się rozpocząć rozważania od wyjaśnienia terminu system. Słownik języka polskiego definiuje go jako:

- układ elementów mających określoną strukturę i stanowiący logicznie uporządkowaną całość;
- zespół wielu urządzeń, dróg, przewodów, itp., funkcjonujących jako całość;
- narządy lub inne części żywego organizmu pełniące razem określone funkcje;
- uporządkowany zbiór twierdzeń, poglądów, tworzących jakąś teorię;
- określony sposób wykonywania jakiejś czynności lub zasady organizacji czegoś;
- forma ustroju państwowego;
- zespół skał powstałych w ciągu jednego okresu geologicznego;

[1] J. Prońko, B. Kaczmarczyk, Bezpieczeństwo państwa – pojęcie, istota, strategia, system, „Rocznik Bezpieczeństwa Morskiego", Rok IX, 2015, s. 117.

- całościowy i uporządkowany zespół zadań połączonych ze sobą stosunkami logicznego wynikania[2].

Bezpieczeństwo narodowe rozpatrywane w różnych kontekstach jest wartością i potrzebą narodową, celem działalności państwa i całego narodu. Postrzegamy je jednak przede wszystkim jako wytwór o charakterze organizacyjnym lub jako system społeczny[3]. Jest to jeden z kluczowych filarów funkcjonowania współczesnego państwa. W warunkach dynamicznych przemian międzynarodowych, rozwoju nowych technologii oraz wzrostu liczby zagrożeń o charakterze hybrydowym i asymetrycznym, zagadnienie to nabiera szczególnego znaczenia. Jest nie tylko domeną wojskowości i działań sił zbrojnych, lecz także przestrzenią aktywności wielu instytucji cywilnych, służb specjalnych, administracji publicznej, a nawet sektora prywatnego i organizacji społecznych.

Bezpieczeństwo narodowe to stan, który umożliwia społeczeństwu oraz państwu rozwój i funkcjonowanie w warunkach stabilności, bez poczucia zagrożenia dla jego wartości podstawowych, takich jak niepodległość, integralność terytorialna, suwerenność, ład konstytucyjny, porządek publiczny czy dobrobyt społeczny. Jest to pojęcie złożone, dynamiczne i trudne do jednoznacznego zdefiniowania, a jego zakres zmienia się wraz z ewolucją zagrożeń i oczekiwań społecznych. W literaturze przedmiotu funkcjonuje wiele definicji bezpieczeństwa. Jedną z klasycznych jest ujęcie, w którym bezpieczeństwo oznacza stan braku zagrożeń oraz zdolność do skutecznego przeciwdziałania im w przypadku wystąpienia[4]. W kontekście narodowym (państwowym), bezpieczeństwo to zorganizowany system działań mających na celu zapewnienie trwałości bytu państwowego, jego niezależności oraz nienaruszalności podstawowych interesów.

Polskie akty normatywne, takie jak Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej (ostatnia z 2020 r.), definiują bezpieczeństwo narodowe jako „zapewnienie państwu i obywatelom możliwości przetrwania, suwerenności, integralności terytorialnej, oraz swobodnego rozwoju

---

[2] Słownik języka polskiego PWN, Wydawnictwo Naukowe PWN, Warszawa 2007, s. 339.

[3] W. Kitler, System bezpieczeństwa narodowego RP – aspekty prawno-organizacyjne, „Wiedza Obronna", vol. 268, no. 3, 2019, s. 6.

[4] T. Bąk, Z. Ciekanowski, Teorie bezpieczeństwa, PWSTE, Jarosław 2012, s. 14.

w warunkach poszanowania prawa i stabilności międzynarodowej”[5]. W ujęciu funkcjonalnym, bezpieczeństwo narodowe obejmuje:

- bezpieczeństwo militarne,
- bezpieczeństwo wewnętrzne,
- bezpieczeństwo polityczne,
- bezpieczeństwo ekonomiczne,
- bezpieczeństwo społeczne,
- bezpieczeństwo ekologiczne,
- bezpieczeństwo informacyjne i cyberbezpieczeństwo.

Tak szerokie ujęcie odzwierciedla współczesne podejście do problematyki bezpieczeństwa, które wykracza poza tradycyjne ramy obronności i obejmuje niemal wszystkie sfery funkcjonowania państwa. Bezpieczeństwo narodowe jako zjawisko społeczne i polityczne, charakteryzuje się kilkoma fundamentalnymi cechami:

- relatywność – nie ma stanu absolutnego bezpieczeństwa; zawsze istnieje pewien poziom ryzyka i niepewności;
- zmienność – charakter i intensywność zagrożeń zmieniają się w czasie w zależności od kontekstu geopolitycznego i technologicznego.
- subiektywność – poczucie bezpieczeństwa obywateli nie zawsze odpowiada obiektywnej ocenie zagrożeń.
- interdyscyplinarność – bezpieczeństwo jest przedmiotem zainteresowania wielu nauk: politologii, socjologii, prawa, ekonomii, nauk wojskowych, informatyki.
- systemowość – bezpieczeństwo narodowe opiera się na współdziałaniu wielu podmiotów w ramach spójnego systemu.

Z punktu widzenia zakresu oddziaływania, bezpieczeństwo narodowe dzieli się na trzy podstawowe poziomy:

- poziom jednostkowy (osobisty) – zapewnienie obywatelowi ochrony życia, zdrowia, wolności i mienia;
- poziom społeczny – ochrona ładu społecznego, integralności grup społecznych, kultury narodowej i wartości konstytucyjnych;
- poziom państwowy – zabezpieczenie suwerenności, terytorium, ustroju politycznego, interesów strategicznych państwa.

[5] Strategia Bezpieczeństwa Narodowego 2020.

Każdy z tych poziomów jest wzajemnie powiązany i wpływa na pozostałe. Na przykład zagrożenia informacyjne mogą jednocześnie oddziaływać na postawy jednostek, spójność społeczną i procesy decyzyjne w państwie. Z kolei w wymiarze przedmiotowym, bezpieczeństwo narodowe można rozpatrywać w kontekście:

- bezpieczeństwa wewnętrznego, czyli ochrony porządku publicznego, integralności instytucji państwowych i obywatelskich wolności,
- bezpieczeństwa zewnętrznego, obejmującego obronność, dyplomację i współpracę międzynarodową,
- bezpieczeństwa informacyjnego, odnoszącego się do kontroli i ochrony przestrzeni informacyjnej,
- bezpieczeństwa gospodarczego, dotyczącego suwerenności ekonomicznej i stabilności rynku.

W XX i XXI wieku pojęcie bezpieczeństwa ulegało istotnej ewolucji. W okresie zimnej wojny dominował paradygmat militarny, koncentrujący się na zagrożeniu ze strony innego państwa. Po roku 1989 nastąpiła zmiana – pojawiły się nowe zagrożenia: terroryzm, przestępczość transgraniczna, migracje, katastrofy naturalne, epidemie, wojna informacyjna. Coraz częściej mówi się też o bezpieczeństwie „zintegrowanym", czyli uwzględniającym współzależność różnych obszarów funkcjonowania państwa i społeczeństwa. Współcześnie bezpieczeństwo traktowane jest jako proces, a nie statyczny stan. Z jednej strony oznacza to konieczność ciągłego monitorowania zagrożeń, z drugiej – budowania odporności systemu państwowego i społecznego na potencjalne kryzysy.

Elementy systemu bezpieczeństwa państwa mogą być rozpatrywane w ujęciu:

- podmiotowym - ogół podmiotów wykonujących określone funkcje (obowiązki) związane z zapewnieniem warunków bezpiecznego funkcjonowania struktur państwa oraz ochroną ludności i majątku narodowego przed skutkami oddziaływań kryzysowych i wojennych;
- strukturalnym - ogół podmiotów wykonujących określone funkcje dotyczące zapewnienia warunków bezpiecznego funkcjonowania struktur państwa oraz ochroną ludności i majątku narodowego przed skutkami oddziaływań kryzysowych i wojennych na wszystkich szczeblach organizacyjnych państwa (poziom krajowy, wojewódzki, powiatowy;

- przedmiotowym - działalność, której przedmiotem jest chronione dobro;
- formalnym (prawnym) - elementy systemu stanowią wyodrębnioną całość, posiadają określoną właściwość, działają w imieniu państwa, działają w zakresie przyznanych im kompetencji[6].

Obecnie jednym z najdynamiczniej rozwijających się obszarów zagrożeń jest przestrzeń informacyjna. Wzrost znaczenia mediów masowych, internetu, platform społecznościowych oraz treści audiowizualnych sprawia, że bezpieczeństwo narodowe coraz silniej powiązane jest z kontrolą i ochroną informacji. Informacja jest jednym z głównych zasobów niematerialnych, a jej bezpieczeństwo dotyczy ochrony przed bezprawną ingerencją i zakłóceniami wytwarzania, przetwarzania, przechowywania i przesyłania wszelkich informacji[7]. Manipulacja informacją, czyli również przekazem medialnym, dezinformacja, propaganda czy tzw. fake newsy stają się narzędziem walki politycznej, ideologicznej, a nawet militarnej. Analiza treści audiowizualnych jako forma monitoringu i przeciwdziałania wpływom destabilizującym, stanowi dziś ważny element strategii bezpieczeństwa państwa. Wymaga to jednak odpowiednich narzędzi, kompetencji oraz zintegrowanych struktur analitycznych, o czym szerzej będzie mowa w dalszych rozdziałach.

Bezpieczeństwo narodowe to pojęcie wielowymiarowe i dynamiczne, które odzwierciedla aktualne wyzwania i potrzeby państwa oraz społeczeństwa. Nie ogranicza się już jedynie do obronności, ale obejmuje także bezpieczeństwo wewnętrzne, informacyjne, społeczne czy ekonomiczne. Rozpoznanie i zrozumienie istoty bezpieczeństwa narodowego stanowi punkt wyjścia dla dalszych rozważań nad rolą analizy treści audiowizualnych, które są współcześnie jednym z kluczowych źródeł wpływu i zagrożeń w przestrzeni informacyjnej.

---

[6] B. Wiśniewski B., System bezpieczeństwa państwa. Konteksty teoretyczne i praktyczne, WSPol, Szczytno 2013, s. 122.

[7] B. Zdrodowski, Istota bezpieczeństwa państwa, „Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate", 9(3), 2019, s. 65.

1.2. Struktura i główne podmioty systemu bezpieczeństwa narodowego

System bezpieczeństwa narodowego Rzeczypospolitej Polskiej to złożona i wielowarstwowa struktura instytucjonalna, której podstawowym celem jest zapewnienie suwerenności państwa, integralności jego terytorium oraz bezpieczeństwa obywateli. Współczesne podejście do tej problematyki zakłada, że skuteczna ochrona państwa przed zagrożeniami wymaga zaangażowania zarówno organów państwowych, jak i społeczeństwa obywatelskiego, a także elastycznego reagowania na zmienne warunki środowiska bezpieczeństwa. W systemie tym funkcjonują instytucje konstytucyjne, administracja rządowa, siły zbrojne, służby specjalne, formacje mundurowe odpowiedzialne za bezpieczeństwo wewnętrzne, a także podmioty pozarządowe i społeczne, które współdziałają w celu zapewnienia spójnej polityki bezpieczeństwa narodowego.

Bezpieczeństwo narodowe w ujęciu strukturalnym (systemowym) odnosi się do kompleksowego przygotowania oraz organizacji państwa w celu zapewnienia bezpieczeństwa w różnych wymiarach. Obejmuje ono działania w sferze prawa i polityki, strategii i obronności, zarówno w aspekcie militarnym, jak i cywilnym, a także aktywność dyplomatyczną, edukacyjną, współpracę międzynarodową oraz procesy integracyjne[8]. System kierowania bezpieczeństwem narodowym, będący integralnym elementem systemu bezpieczeństwa państwa, obejmuje te struktury, których zadaniem jest organizacja i nadzór nad jego funkcjonowaniem[9]. W skład tego systemu wchodzą przede wszystkim organy władzy publicznej oraz osoby kierujące jednostkami realizującymi zadania związane z zapewnieniem bezpieczeństwa narodowego. Dotyczy to również struktur dowodzenia Sił Zbrojnych RP, a także organów doradczych i zaplecza administracyjno-sztabowego. Istotną częścią tego systemu są również określone procedury działania oraz infrastruktura techniczna – takie jak stanowiska dowodzenia, centra zarządzania i systemy łączności – które umożliwiają skuteczne kierowanie bezpieczeństwem państwa.

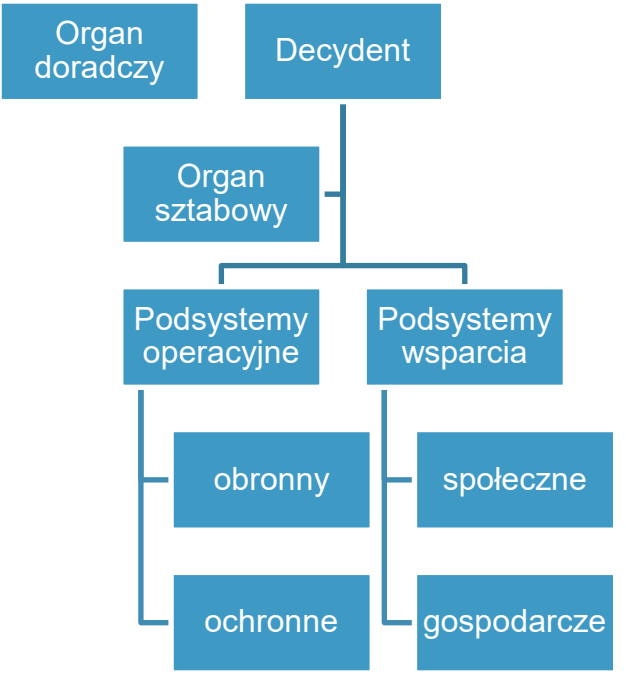
---

[8] W. Pokruszyński, Współczesne bezpieczeństwo narodowe, WSGE, Józefów 2009, s. 179.

[9] J. Prońko, B. Kaczmarczyk, Bezpieczeństwo państwa – pojęcie, istota, strategia, system…, op. cit., s. 121.

W ramach podsystemu kierowania bezpieczeństwem narodowym wyróżnia się kluczowe instytucje, takie jak Parlament, Prezydent RP oraz Rada Ministrów, które pełnią szczególną rolę w procesie podejmowania decyzji strategicznych. Zadaniem całego systemu jest m.in. identyfikowanie i monitorowanie potencjalnych zagrożeń dla bezpieczeństwa kraju, przeciwdziałanie ich powstawaniu zarówno w kraju, jak i poza jego granicami, a także minimalizowanie skutków i usuwanie następstw tych zagrożeń.

Z kolei podsystemy wykonawcze bezpieczeństwa narodowego to siły i zasoby przeznaczone do realizacji zadań wynikających z obowiązujących przepisów prawa. Działają one pod nadzorem organów kierujących bezpieczeństwem i można je podzielić na dwie zasadnicze grupy. Pierwszą stanowią podsystemy operacyjne, obejmujące m.in. komponenty obronne odpowiedzialne za bezpieczeństwo militarne, jak również podsystemy ochronne zajmujące się bezpieczeństwem cywilnym i pozamilitarnym, w tym ochroną ludności. Drugą kategorią są podsystemy wsparcia – zarówno społeczne, jak i gospodarcze – które wspierają funkcjonowanie całego systemu bezpieczeństwa poprzez dostarczanie zasobów, logistyki, wiedzy i wsparcia organizacyjnego. Skonstruowany w ten sposób uniwersalny model systemu bezpieczeństwa narodowego przedstawiono na poniższym rysunku.



Rysunek 1. Uniwersalny model systemu bezpieczeństwa narodowego

Źródło: S. Koziej, System obronny państwa. Cz. 3. Strategia preparacyjna.

Kluczową rolę w systemie bezpieczeństwa odgrywają organy konstytucyjne. Prezydent Rzeczypospolitej Polskiej jako najwyższy przedstawiciel państwa i Zwierzchnik Sił Zbrojnych, posiada szczególne uprawnienia w zakresie bezpieczeństwa[10]. Może podejmować decyzje o użyciu Sił Zbrojnych w obronie kraju, wprowadzać stany nadzwyczajne na wniosek Rady Ministrów oraz przewodniczyć Radzie Bezpieczeństwa Narodowego, która pełni funkcję doradczą w zakresie strategii bezpieczeństwa[11]. Rada Ministrów, z kolei, odpowiada za bieżące zarządzanie bezpieczeństwem państwa oraz koordynację działań poszczególnych ministerstw i organów wykonawczych. Premier i ministrowie realizują politykę bezpieczeństwa na poziomie krajowym oraz odpowiadają za reagowanie w sytuacjach kryzysowych, takich jak katastrofy naturalne, zagrożenia militarne czy ataki cybernetyczne.

[10] A. Mróz-Jagiełło, Rola urzędu prezydenta Rzeczypospolitej Polskiej w systemie bezpieczeństwa narodowego, Problemy Techniki Uzbrojenia, 167(5), 80.

[11] Biała Księga Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej, BBN, Warszawa 2013, s. 250.

Minister Obrony Narodowej kieruje polityką obronną państwa, nadzoruje Siły Zbrojne, planuje działania wojskowe i odpowiada za współpracę w ramach NATO. Minister Spraw Wewnętrznych i Administracji nadzoruje formacje odpowiedzialne za bezpieczeństwo wewnętrzne, takie jak Policja, Straż Pożarna, Straż Graniczna, a także struktury zarządzania kryzysowego.

W strukturze systemu bezpieczeństwa narodowego istotne miejsce zajmują Siły Zbrojne Rzeczypospolitej Polskiej. Ich głównym zadaniem jest ochrona państwa przed zagrożeniami zewnętrznymi, obrona terytorium oraz udział w międzynarodowych operacjach wojskowych i pokojowych. W skład Sił Zbrojnych wchodzą Wojska Lądowe, Siły Powietrzne, Marynarka Wojenna oraz Wojska Obrony Terytorialnej. Szczególnie te ostatnie jako komponent nowego podejścia do obrony powszechnej, odgrywają ważną rolę w budowaniu odporności społecznej, wsparciu administracji publicznej i działaniach w sytuacjach kryzysowych[12]. Wojsko uczestniczy także w ćwiczeniach obronnych, działaniach ratowniczych i ewakuacyjnych, zabezpieczeniu granic i wsparciu służb cywilnych. Współpraca z sojusznikami w ramach NATO oraz Unii Europejskiej wzmacnia potencjał obronny Polski i wpisuje działania wojska w szerszy kontekst międzynarodowego bezpieczeństwa.

Kolejną grupę podmiotów stanowią służby specjalne i instytucje ochrony porządku publicznego. Agencja Bezpieczeństwa Wewnętrznego zajmuje się ochroną państwa przed zagrożeniami wewnętrznymi, w tym zwalczaniem terroryzmu, szpiegostwa, sabotażu i działalności ekstremistycznej[13]. Agencja Wywiadu koncentruje się na działaniach poza granicami kraju, pozyskując informacje istotne dla bezpieczeństwa państwa. Służba Kontrwywiadu Wojskowego oraz Służba Wywiadu Wojskowego zabezpieczają interesy obronne Polski, prowadząc działania kontrwywiadowcze i wywiadowcze w strukturach wojskowych. Centralne Biuro Antykorupcyjne odpowiada za przeciwdziałanie korupcji, szczególnie w obszarach strategicznych dla funkcjonowania państwa, takich jak gospodarka czy administracja publiczna. Formacje takie jak Policja, Państwowa Straż Pożarna oraz

[12] W. Śmiałek (red.), Siły zbrojne w polityce i systemie bezpieczeństwa państwa, Wydawnictwo Naukowe FNCE, Poznań 2023, s. 10.

[13] P. Swoboda, Agencja Bezpieczeństwa Wewnętrznego w systemie bezpieczeństwa państwa, [w:] M. Śliwa, A. Żebrowski, R. Kłaczyński (red.), Bezpieczeństwo RP. Wczoraj i dziś. Studia z zakresu bezpieczeństwa państwa, Wydawnictwo Naukowe Uniwersytetu Pedagogicznego, Kraków 2014, s. 315.

Straż Graniczna odgrywają kluczową rolę w zapewnieniu bezpieczeństwa wewnętrznego, porządku publicznego, ochrony granic i reagowania na sytuacje nadzwyczajne. Policja odpowiada za zwalczanie przestępczości, ochronę osób i mienia, a także za działania prewencyjne. Państwowa Straż Pożarna i ochotnicze straże pożarne wspierają działania ratownicze, gaśnicze i ewakuacyjne. Straż Graniczna zapewnia kontrolę i ochronę granicy państwowej, a także zwalcza nielegalną migrację i przestępczość transgraniczną.

Współczesne podejście do bezpieczeństwa narodowego zakłada również aktywną rolę społeczeństwa obywatelskiego oraz sektora pozarządowego. Organizacje społeczne, fundacje i stowarzyszenia coraz częściej angażują się w działania na rzecz bezpieczeństwa, oferując wsparcie w edukacji, prewencji, analizie zagrożeń czy pomocy humanitarnej. Ich aktywność przyczynia się do zwiększenia świadomości społecznej i budowania kultury bezpieczeństwa. Szczególnie istotne jest zaangażowanie organizacji działających w obszarze edukacji obronnej, ratownictwa, pierwszej pomocy, obrony cywilnej, a także przeciwdziałania dezinformacji i manipulacjom informacyjnym. Społeczeństwo obywatelskie pełni funkcję nie tylko wspierającą, ale i kontrolną wobec instytucji państwowych, co sprzyja przejrzystości i skuteczności podejmowanych działań. Również media jako czwarta władza, mają potencjał do kształtowania opinii publicznej, informowania o zagrożeniach i przeciwdziałania panice.

System bezpieczeństwa narodowego nie funkcjonuje w próżni – jego efektywność zależy od skoordynowanego działania wszystkich wymienionych podmiotów, odpowiedniego przepływu informacji, jasno określonych kompetencji i procedur, a także elastyczności w reagowaniu na nowe typy zagrożeń. Integracja działań różnych instytucji, w tym także samorządów i sektora prywatnego, jest warunkiem niezbędnym do zapewnienia bezpieczeństwa w sytuacjach kryzysowych i długofalowej stabilności państwa. Nowoczesny system bezpieczeństwa to nie tylko obrona przed zagrożeniami, ale również zarządzanie ryzykiem, prewencja, budowanie odporności społecznej oraz adaptacja do zmieniającego się środowiska geopolitycznego i technologicznego.

1.3. Mechanizmy zarządzania bezpieczeństwem państwa - strategiczne dokumenty i polityki

Mechanizmy zarządzania bezpieczeństwem państwa stanowią fundamentalny element organizacji systemu bezpieczeństwa narodowego i są odpowiedzialne za planowanie, koordynację, realizację oraz monitorowanie działań ukierunkowanych na przeciwdziałanie zagrożeniom oraz minimalizowanie ich skutków. W warunkach współczesnych państwo nie może funkcjonować bez sprawnych, zintegrowanych procedur oraz struktur, które umożliwiają zarówno szybkie reagowanie w sytuacjach nadzwyczajnych, jak i długofalowe kształtowanie polityki bezpieczeństwa. Skuteczne zarządzanie bezpieczeństwem wymaga zarówno wypracowania strategicznych dokumentów określających kierunki działań państwa, jak i sprawnego systemu zarządzania kryzysowego oraz mechanizmów wymiany informacji i koordynacji międzyinstytucjonalnej.

Podstawowym elementem mechanizmu zarządzania bezpieczeństwem państwa są dokumenty strategiczne, które stanowią formalny wyraz polityki bezpieczeństwa narodowego i służą jako narzędzie planowania, podejmowania decyzji oraz wdrażania działań operacyjnych i administracyjnych. Kluczowym dokumentem w tym zakresie jest Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej, która definiuje priorytety i cele strategiczne państwa w dziedzinie bezpieczeństwa[14]. Strategia ta określa m.in. zagrożenia i wyzwania dla bezpieczeństwa narodowego, kierunki polityki obronnej, podstawowe zasady reagowania państwa na sytuacje kryzysowe, a także sposób realizacji interesów narodowych w środowisku międzynarodowym. Strategia Bezpieczeństwa Narodowego jest dokumentem nadrzędnym wobec innych planów i programów w zakresie bezpieczeństwa i obronności, a jej aktualizacja odzwierciedla zmieniające się realia geopolityczne, militarne i społeczne. Uzupełnieniem Strategii są inne dokumenty, takie jak:

- Krajowy Plan Zarządzania Kryzysowego,
- Narodowy Program Infrastruktury Krytycznej,
- Strategia Cyberbezpieczeństwa RP,

[14] T. Goryca, Polityka i strategia bezpieczeństwa narodowego Rzeczypospolitej Polskiej – wstęp do dyskusji, Zeszyty Naukowe Państwowej Wyższej Szkoły Zawodowej im. Witelona w Legnicy, 41(4), 43-44.

- inne strategie sektorowe (np. w zakresie ochrony infrastruktury krytycznej, ochrony ludności, bezpieczeństwa energetycznego, ochrony środowiska). Dokumenty te określają zadania dla poszczególnych resortów i instytucji, definiują procedury reagowania, a także tworzą podstawy do planowania budżetowego i organizacyjnego.

Obok planowania strategicznego, kluczową rolę w mechanizmach zarządzania bezpieczeństwem odgrywa system reagowania kryzysowego. Jest to zintegrowany układ działań, procedur i struktur organizacyjnych mających na celu zapewnienie szybkiej i skutecznej reakcji państwa w przypadku wystąpienia zagrożeń o charakterze nagłym, masowym lub nadzwyczajnym. Reagowanie kryzysowe obejmuje wszystkie etapy procesu zarządzania kryzysem: od zapobiegania, poprzez przygotowanie, reagowanie, aż po odbudowę po kryzysie.

W Polsce system zarządzania kryzysowego opiera się na przepisach ustawy z dnia 26 kwietnia 2007 r. o zarządzaniu kryzysowym, która definiuje zasady funkcjonowania administracji publicznej w sytuacjach zagrożenia[15]. System ten działa na czterech poziomach: krajowym, wojewódzkim, powiatowym i gminnym. Na każdym z tych poziomów funkcjonują centra zarządzania kryzysowego, których zadaniem jest gromadzenie i analiza informacji o zagrożeniach, koordynowanie działań służb ratowniczych i porządkowych, a także współpraca z podmiotami pozarządowymi i sektorem prywatnym. Szczególną rolę w krajowym systemie pełni Rządowe Centrum Bezpieczeństwa (RCB), które prowadzi całodobowy monitoring zagrożeń, wspiera organy administracji w podejmowaniu decyzji, zarządza systemem ostrzegania ludności (Alert RCB), a także koordynuje działania w przypadku wprowadzenia stanów nadzwyczajnych. RCB odpowiada także za funkcjonowanie rządowego zespołu zarządzania kryzysowego, który podejmuje decyzje operacyjne na najwyższym szczeblu. System zarządzania kryzysowego uwzględnia także integrację z planowaniem obronnym, systemem obrony cywilnej oraz międzynarodowymi mechanizmami pomocy i współpracy.

Zarządzanie bezpieczeństwem państwa nie może być skuteczne bez sprawnie działającego mechanizmu koordynacji międzyinstytucjonalnej i wymiany informacji. Współczesne zagrożenia, takie jak ataki

[15] K. Piękoś, Organizacja systemu zarządzania kryzysowego w Polsce, Annales Scientia Politica, 2, 44.

cybernetyczne, katastrofy naturalne, pandemie, masowe migracje czy działania hybrydowe, wymagają ścisłej współpracy wielu podmiotów – zarówno na szczeblu rządowym, jak i samorządowym, a także pomiędzy służbami mundurowymi, wojskiem, służbami specjalnymi, administracją cywilną, organizacjami społecznymi oraz sektorem prywatnym. Koordynacja ta obejmuje zarówno wymianę danych o zagrożeniach, jak i wspólne planowanie działań, synchronizację zasobów, ustalanie priorytetów oraz współdziałanie operacyjne.

W Polsce za koordynację działań w sytuacjach nadzwyczajnych odpowiadają organy administracji rządowej i samorządowej, które działają na podstawie ustaw i strategii oraz są wspierane przez specjalistyczne zespoły i centra operacyjne. Przykładem może być Stały Zespół do spraw Zagrożeń Terrorystycznych, funkcjonujący przy Ministrze Spraw Wewnętrznych i Administracji, czy Zespół do spraw Cyberbezpieczeństwa przy Prezesie Rady Ministrów. Równie istotna jest rola Krajowego Systemu Wykrywania Skażeń i Alarmowania oraz systemu ostrzegania ludności, które działają w oparciu o współdziałanie wielu instytucji – od wojska, przez służby ratownicze, aż po administrację lokalną. Wymiana informacji wspierana jest przez specjalistyczne platformy komunikacji, w tym systemy łączności kryzysowej, centra analityczne oraz bazy danych zagrożeń. Szczególną rolę odgrywa tutaj także współpraca międzynarodowa – Polska jako członek NATO, UE i ONZ uczestniczy w systemach wymiany informacji, ćwiczeniach oraz wspólnych działaniach prewencyjnych i reagujących.

Zarządzanie bezpieczeństwem to również obszar wymagający ciągłego doskonalenia. Wyzwania takie jak postępująca cyfryzacja, rozwój sztucznej inteligencji, globalne zmiany klimatyczne czy wzrost znaczenia zagrożeń hybrydowych sprawiają, że państwo musi elastycznie dostosowywać swoje struktury i procedury do nowych realiów. Kluczowe staje się nie tylko reagowanie na incydenty, ale również inwestowanie w systemy predykcji, analitykę zagrożeń oraz rozwój kompetencji kadry zarządzającej. Wymaga to również zaangażowania środowisk akademickich i eksperckich, które dostarczają wiedzy o charakterze strategicznym, a także integracji działań edukacyjnych i informacyjnych na poziomie społecznym.

Efektywne mechanizmy zarządzania bezpieczeństwem państwa muszą także obejmować odpowiednie przygotowanie infrastruktury technicznej – zarówno fizycznej, jak i cyfrowej. Stanowiska dowodzenia, centra

kierowania, mobilne centra zarządzania kryzysowego, systemy łączności rezerwowej, zasoby logistyczne i medyczne to niezbędne komponenty, które decydują o sprawności działania całego systemu w sytuacjach kryzysowych. Infrastruktura ta musi być odporna na zakłócenia, ataki i awarie, a jednocześnie dostosowana do realiów funkcjonowania w środowisku wielodomenowym – lądowym, powietrznym, morskim, kosmicznym i cyfrowym. Odpowiednie planowanie i utrzymanie infrastruktury bezpieczeństwa to nie tylko zadanie dla państwa, ale również wyzwanie dla partnerów prywatnych i sektora technologicznego, którzy coraz częściej zarządzają infrastrukturą krytyczną, taką jak sieci energetyczne, telekomunikacyjne, transportowe czy informatyczne.

Współczesne mechanizmy zarządzania bezpieczeństwem państwa muszą opierać się na zasadzie kompleksowości, elastyczności i współpracy. Żaden z podmiotów – nawet najbardziej wyspecjalizowany – nie jest w stanie samodzielnie sprostać skali i złożoności zagrożeń, jakie występują w XXI wieku. Kluczowe jest więc integrowanie działań, budowanie odporności systemu, efektywne planowanie i ciągła modernizacja struktur. Tylko wówczas państwo może skutecznie chronić swoich obywateli, terytorium i interesy narodowe, niezależnie od tego, czy zagrożeniem będzie atak militarny, powódź, cyberatak, czy też dezinformacja.

1.4. Informacja i komunikacja jako element bezpieczeństwa

Współczesne systemy bezpieczeństwa narodowego nie mogą być skutecznie realizowane bez uwzględnienia znaczenia informacji i komunikacji. Informacja – zarówno w ujęciu strategicznym, operacyjnym, jak i społecznym – stanowi jeden z kluczowych zasobów państwa, którego właściwe zarządzanie może decydować o skuteczności reagowania na zagrożenia, efektywności podejmowanych działań oraz stabilności systemu politycznego i społecznego. Syntetycznego przeglądu definicji pojęcia informacji dokonał w literaturze przedmiotu Krzysztof Liedel, którego zdaniem podejmowane przez poszczególnych badaczy próby definicyjne nie przyniosły zadowalających rezultatów, bowiem dotykają one jedynie wybranych aspektów informacji. Przywołany autor przytacza następujące definicje:

- informacja to zbiór faktów, zdarzeń, cech itp. określonych obiektów (rzeczy, procesów, systemów) zawarty w wiadomości (komunikacie), ujęty i podany w takiej postaci (formie), że pozwala odbiorcy ustosunkować się do zaistniałej sytuacji i podjąć odpowiednie działania umysłowe lub fizyczne (Piotr Sienkiewicz);
- informacja jest transformacją jednego komunikatu asocjacji informacyjnej w drugi komunikat tej asocjacji; informowanie jest to transformowanie informacji zawartych w łańcuchu oryginałów w informacje zawarte w łańcuchu obrazów (Marian Mazur);
- informacja to wiedza przekazywana przez innych ludzi bądź uzyskiwana przez studia, obserwacje i badania (Andrew Webster);
- informacja to czynnik sterujący strumieniami zasileń, wykorzystywany w organizmach żywych lub maszynach do bardziej sprawnego, efektywnego i celowego działania (Edward Kowalczyk);
- informacja – bodziec oddziaływujący na układ recepcyjny człowieka, powodujący wytwarzanie w jego wyobraźni przedmiotu myślowego, odzwierciedlającego obraz rzeczy materialnej lub abstrakcyjnej, który w jego przekonaniu (świadomości) kojarzy się jakoś z tym bodźcem; oznacza to, że informacje to nie tylko te doznania, które inspirują umysł ludzki do pewnej wyobraźni; jej istnienie jest relatywnie związane z istnieniem człowieka i jego umysłem (Leopold Ciborowski);
- informacją jest nie tylko wiadomość o czymś, ale także każda decyzja, sugestia czy polecenie (Norbert Wiener);
- informacja to wiadomość uzyskana przez człowieka poprzez obserwację lub czynność umysłową, podlegającą przekazowi w układzie nadawca – odbiorca (Henryk Greniewski)[16].

Komunikacja z kolei jako proces przekazywania, przetwarzania i odbioru informacji, umożliwia koordynację działań, budowanie świadomości sytuacyjnej oraz kształtowanie opinii publicznej[17]. W dobie społeczeństwa informacyjnego i błyskawicznego przepływu danych, informacja stała się równie istotnym elementem bezpieczeństwa jak tradycyjne środki

[16] K. Liedel, Zarządzanie informacją w walce z terroryzmem, Wydawnictwo Trio, Warszawa 2010, s. 44.

[17] J. Nowicka, W. Załoga, Z. Ciekanowski, Komunikacja strategiczna w naukach o zarządzaniu i jakości oraz w naukach o bezpieczeństwie, Zeszyty Naukowe Wyższej Szkoły Zarządzania i Ochroną Pracy w Katowicach, 1(14), 2018, 194.

obronne. W celu usystematyzowania pojęć, które są w tym kontekście używane, należy zauważyć, że w ramach kultury bezpieczeństwa wymienia się komponent informacyjny. Tworzą go następujące elementy:

- kultura informacyjna – szczególnie istotna w kontekście kształtowania świadomości informacyjnej społeczeństwa;
- ekologia informacji – promocja zrównoważonego rozwoju człowieka w środowisku technicznym (wirtualnym), jak i w świecie realnym;
- wychowanie do informacji – wspomaganie rozwoju w kontekście korzystania z informacji, co ma ułatwiać dokonywanie racjonalnych wyborów;
- kultura organizacyjna – dbanie o bezpieczeństwo informacji poprzez zapewnienie odpowiednich modeli zachowań;
- polityka informacyjna – jej cele to utrzymanie bezpieczeństwa informacyjnego w wymiarze społecznym;
- kultura bezpieczeństwa informacyjnego – ma umożliwiać budowanie odporności na ataki informacyjne oraz medialne manipulacje[18].

W nowoczesnych państwach informacja odgrywa podwójną rolę – z jednej strony jest narzędziem zarządzania i podejmowania decyzji, z drugiej stanowi potencjalny wektor zagrożeń[19]. Systemy bezpieczeństwa muszą zatem nie tylko zapewniać efektywny obieg informacji wewnętrznej – między instytucjami, służbami, a także między państwem a obywatelem – ale również chronić przestrzeń informacyjną przed manipulacją, dezinformacją i wrogim wpływem. Kluczowe znaczenie ma tu zapewnienie integralności, dostępności i poufności informacji, które mogą mieć wpływ na funkcjonowanie państwa oraz bezpieczeństwo obywateli. Informacja o charakterze strategicznym, operacyjnym, wywiadowczym, technologicznym, ale też społeczno-kulturowym, wymaga nie tylko analizy, lecz także odpowiedniego zabezpieczenia.

Przestrzeń informacyjna we współczesnym świecie stała się pełnoprawnym środowiskiem działań strategicznych. Obok lądu, morza, powietrza i cyberprzestrzeni, właśnie przestrzeń informacyjna stanowi dziś pole rywalizacji państw, grup interesów, organizacji międzynarodowych, mediów

[18] H. Batorowska, Kultura bezpieczeństwa informacyjnego w środowisku walki o przewagę informacyjną, Wydawnictwo Libron, Kraków 2021, s. 260-265.

[19] T. R. Aleksandrowicz, Bezpieczeństwo informacyjne państwa, Studia Politologiczne, 49, 48.

i podmiotów prywatnych[20]. To w niej toczą się zmagania o wpływy, kontrolę nad opinią publiczną, kreowanie narracji oraz kształtowanie zachowań społecznych. Zmieniające się technologie komunikacyjne – przede wszystkim rozwój internetu, mediów społecznościowych, transmisji audiowizualnej w czasie rzeczywistym – sprawiły, że przestrzeń informacyjna stała się niezwykle dynamiczna, podatna na zakłócenia, a zarazem trudna do kontrolowania. Państwa, które nie są w stanie skutecznie monitorować, analizować i kształtować własnego środowiska informacyjnego, narażone są na destabilizację, spadek zaufania społecznego, a w skrajnych przypadkach – na utratę zdolności do zarządzania kryzysami.

Współczesne zagrożenia w przestrzeni informacyjnej przybierają różne formy – od subtelnych działań perswazyjnych po skoordynowane kampanie dezinformacyjne. Dezinformacja rozumiana jako celowe rozpowszechnianie fałszywych lub zmanipulowanych informacji w celu wywarcia określonego wpływu na społeczeństwo, instytucje lub procesy polityczne, stanowi dziś jedno z najpoważniejszych wyzwań dla bezpieczeństwa państw demokratycznych. Jej celem może być m.in. wzbudzenie paniki, podważenie autorytetu władz, wywołanie chaosu informacyjnego, a także osłabienie zaufania społecznego do instytucji publicznych. Dezinformacja często przybiera formę tzw. operacji wpływu, w których przekaz medialny jest wykorzystywany jako narzędzie realizacji interesów politycznych lub strategicznych innych państw[21]. Szczególnie niebezpieczne są działania prowadzone przez podmioty państwowe lub wspierane przez państwa, które za pośrednictwem kanałów informacyjnych próbują ingerować w procesy decyzyjne, wybory, debaty publiczne czy kierunek polityki zagranicznej.

Propaganda jako forma komunikacji perswazyjnej, mająca na celu kształtowanie postaw i przekonań odbiorców w określonym kierunku, również stanowi narzędzie oddziaływania w przestrzeni informacyjnej[22]. Choć propaganda była wykorzystywana w działaniach politycznych i wojennych od wieków, to współczesne technologie nadały jej nowy wymiar. Dzięki analizie danych, mikrotargetowaniu, psychometrii i sztucznej inteligencji,

[20] D. Kaźmierczak, Walka informacyjna we współczesnych konfliktach i jej społeczne konsekwencje, Annales Universitatis Paedagogicae Cracoviensis. 250, Studia de Securitate et Educatione Civili, 7, 127.

[21] T. Kacała, Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa, Przegląd Prawa Konstytucyjnego, 2(24), 54.

[22] M. J. Wachowicz, Ujęcie teoretyczne pojęcia dezinformacji, Wiedza Obronna, 1-2, 239.

możliwe jest dziś precyzyjne docieranie z określonym przekazem do wybranych grup społecznych, a nawet do pojedynczych jednostek. Dodatkowo należy podkreślić, że ani dezinformacja, ani propaganda nie są celem, a środkiem do osiągnięcia określonego, z reguły długofalowego, celu politycznego lub wojskowego[23]. Natomiast manipulacja informacyjna obejmująca zarówno dezinformację, jak i propagandę często posługuje się także tzw. deepfake'ami, zmanipulowanymi obrazami i dźwiękiem, które mogą tworzyć fałszywe relacje z wydarzeń, wypowiedzi publiczne czy dowody wizualne.

Wszystko to sprawia, że treści medialne zwłaszcza audiowizualne stają się jednym z głównych obszarów zainteresowania służb bezpieczeństwa, analityków, ekspertów i instytucji publicznych. W przestrzeni informacyjnej toczą się bowiem nie tylko wojny narracyjne, ale też operacje psychologiczne (psyops), których celem jest zmiana zachowań społecznych, osłabienie morale, podsycanie konfliktów wewnętrznych i polaryzacja opinii publicznej. Zdolność państwa do rozpoznawania tego typu zagrożeń, przeciwdziałania im oraz budowania odporności informacyjnej społeczeństwa staje się dziś jednym z kluczowych aspektów bezpieczeństwa narodowego. W tym kontekście szczególnego znaczenia nabiera analiza treści audiowizualnych – jako metoda identyfikowania manipulacji, monitorowania przekazów medialnych, oceny spójności komunikacji strategicznej państwa oraz wykrywania źródeł zagrożeń informacyjnych.

Państwo, które chce skutecznie chronić swoją przestrzeń informacyjną, musi nie tylko tworzyć mechanizmy szybkiego reagowania na fałszywe przekazy, ale też inwestować w edukację medialną obywateli, wzmacniać rolę niezależnych mediów oraz tworzyć systemy wczesnego ostrzegania i przeciwdziałania dezinformacji. Niezwykle ważna jest też koordynacja działań między instytucjami – służbami specjalnymi, resortami odpowiedzialnymi za bezpieczeństwo, administracją rządową i lokalną – a także współpraca z sektorem prywatnym, w szczególności z dostawcami usług cyfrowych, mediami oraz firmami z branży IT i cyberbezpieczeństwa. Wzrost znaczenia informacji i komunikacji w systemie bezpieczeństwa narodowego wiąże się również z koniecznością redefinicji wielu dotychczasowych pojęć i procedur.

[23] R. Bielawski, B. Grenda, P. Majdan, Wieloaspektowa ewaluacja wykorzystania mediów społecznościowych na potrzeby kierowania SBN RP (ewaluacji ryzyka cyberzagrożeń BN), Akademia Sztuki Wojennej, 2017, s. 69-70.

Bezpieczeństwo informacyjne nie jest dziś wyłącznie domeną wojskową czy wywiadowczą – dotyczy całego społeczeństwa, mediów, edukacji, administracji i gospodarki. Wymaga ono nowego podejścia do tworzenia polityki informacyjnej, bardziej zintegrowanego systemu analiz, szkoleń i ćwiczeń, jak również rozbudowy potencjału analitycznego instytucji państwowych. W tym kontekście rośnie znaczenie interdyscyplinarności – łączenia wiedzy z zakresu politologii, socjologii, informatyki, psychologii społecznej i medioznawstwa – która pozwala lepiej zrozumieć mechanizmy wpływu w środowisku informacyjnym.

Informacja i komunikacja to dziś nie tylko element wspierający bezpieczeństwo państwa, ale jeden z jego kluczowych fundamentów. Przestrzeń informacyjna jako nowoczesne pole rywalizacji i konfliktu, wymaga nieustannego monitorowania, analizy i ochrony. Państwa, które nie potrafią skutecznie zarządzać tym obszarem, narażają się na utratę suwerenności informacyjnej, destabilizację systemu politycznego, osłabienie więzi społecznych i zmniejszenie zdolności do reagowania na kryzysy. Z kolei państwa zdolne do prowadzenia aktywnej, świadomej i zintegrowanej polityki informacyjnej zyskują przewagę strategiczną – zarówno w wymiarze wewnętrznym, jak i międzynarodowym.

II

System bezpieczeństwa informacyjnego państwa

2.1. Pojęcie i zakres bezpieczeństwa informacyjnego

Bezpieczeństwo informacyjne stanowi jeden z fundamentalnych komponentów bezpieczeństwa narodowego we współczesnym świecie, który coraz wyraźniej kształtowany jest przez przepływy informacji, dominację mediów cyfrowych i dynamikę przestrzeni komunikacyjnej[24]. Jest to pojęcie złożone i wielowymiarowe, funkcjonujące zarówno w dyskursie prawnym i politycznym, jak i w refleksji akademickiej oraz praktyce instytucjonalnej. Choć brakuje jednej, powszechnie uznanej definicji bezpieczeństwa informacyjnego, to jednak w różnych ujęciach wyraźnie zarysowują się jego kluczowe cechy: ochrona przed zagrożeniami związanymi z przetwarzaniem, przekazywaniem i odbiorem informacji, a także zapewnienie integralności i suwerenności przestrzeni informacyjnej państwa. Terminem tym określa się także wszelkie działania, systemy oraz metody zmierzające do zabezpieczenia zasobów informacyjnych gromadzonych, przetwarzanych, przechowywanych w pamięciach komputerów oraz sieciach teleinformatycznych, a także organizację przepływu informacji pomiędzy organami władzy[25].
W ramach takiego podejścia należy uwzględnić szereg uwarunkowań bezpieczeństwa informacyjnego, a przede wszystkim to, że:

- informacja stanowi zasób strategiczny państwa;

[24] K. Wąsik, Bezpieczeństwo Informacyjne Państwa [w:] 13th Eurasian Conferences on Language and Social Sciences, Daugavpils University, LATVIA, February 5 - 6, 2022, s. 116.

[25] P. Bączek, Zagrożenia informacyjne a bezpieczeństwo państwa polskiego, Wydawnictwo Adam Marszałek, Toruń 2005, s. 71.

- informacja i wynikająca z niej wiedza oraz technologie informatyczne stają się podstawowym czynnikiem wytwórczym;
- szeroko rozumiany sektor informacyjny generuje znaczną część dochodu narodowego;
- procesy decyzyjne w innych sektorach gospodarki i życia społecznego są w znacznej mierze uzależnione od systemów przetwarzania i przesyłania informacji;
- zakłócenie prawidłowości działania systemów informacyjno-sterujących nie wymaga wysokich nakładów materialnych;
- rywalizacja pomiędzy przeciwnikami przeniesie się na płaszczyznę walki informacyjnej[26];
- technologie informatyczne stały się istotnym elementem funkcjonowania sił zbrojnych[27];
- media masowe mogą być wykorzystywane jako narzędzia skutecznego zakłócania informacyjnego, np. na drodze dezinformacji[28]

Bezpieczeństwo informacyjne nie jest terminem zdefiniowanym w przepisach prawa – w polskim porządku prawnym brak jest aktów normatywnych, w których ustawodawca wprost posługuje się tym pojęciem. W konsekwencji nie istnieje również jego definicja legalna. Mimo to należy zauważyć, że termin ten funkcjonuje samodzielnie w ramach nauki o bezpieczeństwie, a także pełni istotną rolę w konstruowaniu koncepcji i rozwiązań o charakterze normatywnym. Pojęcie bezpieczeństwa informacyjnego należy zatem rozpatrywać zarówno z perspektywy materialnoprawnej, jak i proceduralnej. W pierwszym ujęciu stanowi ono dobro prawne, które – z uwagi na swoją specyfikę – wymaga odrębnej ochrony prawnej, niezależnie od innych dóbr chronionych. W aspekcie proceduralnym z kolei bezpieczeństwo informacyjne pełni funkcję dyrektywy działania – stanowi normę kompetencyjną, na podstawie, której określone organy państwowe mogą podejmować działania wpływające na sytuację prawną innych podmiotów, w tym zakresie ich praw i obowiązków dotyczących zapewnienia bezpieczeństwa oraz

[26] K. Liedel, Bezpieczeństwo informacyjne państwa, [w:] Transsektorowe obszary bezpieczeństwa narodowego, K. Liedel (red.), Warszawa 2011, s. 57.

[27] B. Balcerowicz, Siły zbrojne w czasie pokoju, kryzysu i wojny, Wydawnictwo Naukowe Schoolar, Warszawa 2010, s. 219.

[28] K. Liedel, Bezpieczeństwo informacyjne…, op. cit., s. 57-58.

przeciwdziałania zagrożeniom[29]. Odczytywanie w pojęciu „bezpieczeństwo państwa" normy kompetencyjnej przynależnej organom administracji publicznej wynika z faktu, że to właśnie państwo ponosi zasadniczą odpowiedzialność za zapewnienie bezpieczeństwa we wszystkich jego wymiarach[30].

Można przyjąć, iż w ujęciu prawnym bezpieczeństwo informacyjne odnosi się przede wszystkim do zapewnienia ochrony systemów i zasobów informacji przed nieuprawnionym dostępem, naruszeniem poufności, modyfikacją, zniszczeniem lub zakłóceniem. Definicje te często zawierają się w ramach pojęcia bezpieczeństwa teleinformatycznego lub cyberbezpieczeństwa, lecz w praktyce wykraczają poza aspekt techniczny, obejmując także kontekst społeczny i regulacyjny. W dokumentach strategicznych, takich jak Strategia Bezpieczeństwa Narodowego RP czy Strategia Cyberbezpieczeństwa RP, bezpieczeństwo informacyjne rozumiane jest szerzej – jako ochrona społeczeństwa, instytucji i państwa przed szkodliwym wpływem informacji, dezinformacją, propagandą, a także przed utratą zdolności do suwerennego kształtowania własnej polityki informacyjnej.

Z perspektywy strategicznej, bezpieczeństwo informacyjne postrzegane jest jako zdolność państwa do ochrony i kontroli nad obiegiem informacji istotnych dla interesu publicznego oraz bezpieczeństwa narodowego. Obejmuje ono działania w zakresie prewencji, wykrywania, przeciwdziałania i neutralizacji wszelkich form manipulacji informacyjnej, operacji psychologicznych, kampanii dezinformacyjnych i wpływu zewnętrznego. W tym ujęciu bezpieczeństwo informacyjne jest istotnym elementem obrony hybrydowej, której celem jest nie tylko ochrona infrastruktury, ale także tożsamości narodowej, jedności społecznej i zdolności decyzyjnej państwa. W refleksji akademickiej, bezpieczeństwo informacyjne definiowane jest jako stan funkcjonowania systemów społecznych i technicznych, w którym informacja – jako zasób – jest chroniona przed zagrożeniami zewnętrznymi i wewnętrznymi, a jednocześnie możliwe jest jej wykorzystanie zgodnie z interesem

[29] J. Taczkowska-Olszewska, Bezpieczeństwo informacyjne jako kategoria prawna. Ujęcie teoretyczne [w:] Bezpieczeństwo informacyjne. Aspekty prawno-administracyjne, W. Kitler, J. Taczkowska-Olszewska (red.), Towarzystwo Wiedzy Obronnej, Warszawa 2017, s. 47.

[30] Z. Nowakowski, M. Pomykała, J. Rajchel, Pojęcie bezpieczeństwa i porządku publicznego [w:] Administracja bezpieczeństwa i porządku publicznego ze szczególnym uwzględnieniem aspektów prawno-organizacyjnych Policji, K. Rajchel (red.), Towarzystwo Naukowe Powszechne, Warszawa 2009, s. 48

publicznym, normami społecznymi i konstytucyjnymi zasadami demokracji. Szczególny nacisk kładzie się na analizę zagrożeń narracyjnych, kształtowania percepcji i oddziaływania na emocje odbiorców – co ma kluczowe znaczenie w epoce mediów audiowizualnych i społeczeństwa informacyjnego.

Bezpieczeństwo informacyjne bywa niekiedy utożsamiane z cyberbezpieczeństwem, choć w istocie są to pojęcia komplementarne, lecz nie tożsame. Cyberbezpieczeństwo koncentruje się na ochronie zasobów teleinformatycznych – systemów, sieci, danych – przed zakłóceniami, włamaniami czy atakami. Z kolei bezpieczeństwo informacyjne obejmuje również obszar treści – ich prawdziwość, kontekst, intencję nadawcy oraz sposób wpływu na odbiorcę. W tym ujęciu cyberbezpieczeństwo stanowi komponent techniczny szerszego systemu ochrony informacyjnej państwa i społeczeństwa.

Współczesna teoria i praktyka bezpieczeństwa informacyjnego wskazują na jego wielowymiarowy charakter. Wymiar techniczny dotyczy ochrony infrastruktury cyfrowej, systemów komunikacji, baz danych i sieci przesyłu informacji. Wymiar społeczny obejmuje budowanie zaufania do instytucji, odporność społeczeństwa na manipulację, wspieranie pluralizmu medialnego oraz edukację obywateli. Wymiar psychologiczny koncentruje się na percepcji odbiorców, podatności na emocjonalne przekazy, efekt potwierdzenia i panikę informacyjną. Z kolei wymiar kulturowy wiąże się z ochroną tożsamości narodowej, symboli wspólnotowych, języka oraz narracji konstytuujących wspólnotę polityczną. W kontekście międzynarodowym, bezpieczeństwo informacyjne staje się przedmiotem rosnącego zainteresowania organizacji międzyrządowych oraz struktur sojuszniczych. NATO w swoich oficjalnych dokumentach podkreśla, że współczesne operacje hybrydowe obejmują również komponent informacyjny, a odporność państw członkowskich na dezinformację i wojny narracyjne jest warunkiem ich wspólnej obrony[31]. W 2014 r. po kryzysie krymskim Sojusz uruchomił StratCom COE w Rydze – Centrum Doskonałości Komunikacji Strategicznej, którego zadaniem jest m.in. badanie i przeciwdziałanie dezinformacji, wspieranie państw członkowskich w prowadzeniu komunikacji strategicznej oraz analiza działań informacyjnych aktorów państwowych i niepaństwowych. Ponadto NATO uznaje przestrzeń informacyjną za domenę potencjalnych działań operacyjnych na równi z przestrzenią cybernetyczną, morską, lądową i powietrzną.

[31] P. Szymański, NATO i Unia Europejska wobec zagrożeń hybrydowych, Ośrodek Studiów Wschodnich, 328, 1.

Unia Europejska traktuje bezpieczeństwo informacyjne jako jeden z filarów szerszej strategii cyfrowej i bezpieczeństwa wewnętrznego. W ramach Digital Services Act (DSA) oraz Digital Markets Act (DMA) wprowadzono regulacje zobowiązujące platformy cyfrowe do przeciwdziałania dezinformacji i większej przejrzystości algorytmów. Digital Services Act (DSA) oraz Digital Markets Act (DMA) tworzą jeden zbiór przepisów, które mają zastosowanie w całej UE. Mają one 2 główne cele:

- stworzenie bezpieczniejszej przestrzeni cyfrowej, w której chronione są prawa podstawowe wszystkich użytkowników usług cyfrowych;
- stworzenie równych warunków działania w celu wspierania innowacji, wzrostu gospodarczego i konkurencyjności, zarówno na jednolitym rynku europejskim, jak i na całym świecie[32].

UE prowadzi również działania edukacyjne i badawcze w ramach programów Horizon Europe, EDMO (European Digital Media Observatory) oraz inicjatyw takich jak Code of Practice on Disinformation. W 2022 r. zaktualizowano również dokument Strategic Compass, w którym uznano dezinformację i manipulacje narracyjne za poważne zagrożenie dla jedności Unii i jej obywateli.

W podobnym kierunku zmierzają działania ONZ, która w ramach Global Digital Compact oraz inicjatyw UNESCO promuje koncepcję pluralizmu medialnego, edukacji cyfrowej oraz etycznego wykorzystania informacji i technologii[33]. Równolegle wiele państw – m.in. Estonia, Finlandia, Kanada, USA – rozwija własne strategie bezpieczeństwa informacyjnego, łącząc aspekty regulacyjne, technologiczne i edukacyjne.

W tym kontekście informacja jawi się jako zasób o strategicznym znaczeniu dla państwa – porównywalny z zasobami militarnymi, energetycznymi czy technologicznymi. Kontrola nad informacją, jej źródłami, kanałami przepływu i mechanizmami przetwarzania staje się jednym z kluczowych elementów władzy politycznej, ekonomicznej i społecznej. Państwo, które traci zdolność do ochrony własnej przestrzeni informacyjnej, naraża się nie tylko na zakłócenia w funkcjonowaniu instytucji publicznych, ale także na destabilizację opinii publicznej, utratę zaufania społecznego

[32] The Digital Services Act package. Shaping Europe's digital future, https://digitalstrategy.ec.europa.eu/en/policies/digital-services-act-package (dostęp: 19.04.2025)

[33] A. Gwagwa, W. J. T. Mollema, How could the United Nations Global Digital Compact prevent cultural imposition and hermeneutical injustice?. Patterns, 5(11).

i osłabienie spójności narodowej. Dlatego też inwestycje w bezpieczeństwo informacyjne – zarówno na poziomie krajowym, jak i w ramach współpracy międzynarodowej – powinny być traktowane jako integralna część polityki bezpieczeństwa narodowego oraz fundament odporności strategicznej państwa w dobie globalnej rywalizacji narracyjnej.

2.2. Struktura systemu bezpieczeństwa informacyjnego

System bezpieczeństwa informacyjnego państwa jako część szerszego systemu bezpieczeństwa narodowego, ma charakter złożony, dynamiczny i wielowarstwowy. Jego zadaniem jest zapewnienie ochrony informacji i przestrzeni informacyjnej przed wszelkimi zagrożeniami – zarówno technologicznymi, jak i narracyjnymi – które mogą zagrażać stabilności państwa, interesowi publicznemu oraz integralności debaty społecznej. Struktura tego systemu opiera się na czterech podstawowych filarach: prewencji, monitorowaniu, reagowaniu i edukacji. Są one wzajemnie powiązane i powinny być realizowane spójnie na wszystkich poziomach funkcjonowania państwa.

Prewencja stanowi pierwszy, strategiczny poziom działania systemu bezpieczeństwa informacyjnego. Obejmuje tworzenie warunków organizacyjnych, prawnych i technologicznych umożliwiających minimalizację ryzyka wystąpienia zagrożeń informacyjnych. Do działań prewencyjnych należy m.in. wzmacnianie odporności systemów teleinformatycznych, rozwój infrastruktury krytycznej o znaczeniu informacyjnym, przeciwdziałanie kampaniom dezinformacyjnym poprzez edukację społeczeństwa oraz współpracę międzynarodową w zakresie wymiany dobrych praktyk i wczesnego ostrzegania. Prewencja obejmuje również działania instytucjonalne, takie jak tworzenie aktów normatywnych regulujących odpowiedzialność platform cyfrowych czy rozwój polityk komunikacji strategicznej na poziomie państwowym.

Drugim filarem jest **monitorowanie**, rozumiane jako stała analiza przestrzeni informacyjnej w celu identyfikacji potencjalnych zagrożeń. Monitorowanie obejmuje obserwację tradycyjnych i cyfrowych kanałów komunikacji, w tym mediów społecznościowych, forów, portali informacyjnych i komunikatorów. W jego ramach funkcjonują specjalistyczne zespoły analityczne, zarówno w strukturach państwowych (np. RCB, MON, ABW), jak i instytucjach akademickich czy pozarządowych. Efektywne monitorowanie

wymaga również integracji systemów analizy danych, w tym wykorzystania narzędzi sztucznej inteligencji, uczenia maszynowego oraz metod przetwarzania języka naturalnego i obrazu. Jego celem jest nie tylko wykrycie dezinformacji, ale także ocena jej zasięgu, skutków oraz potencjalnych źródeł.

Trzecim komponentem systemu jest **reagowanie**, czyli podejmowanie działań operacyjnych i strategicznych w odpowiedzi na zaistniałe incydenty informacyjne. Może ono obejmować publikację oficjalnych komunikatów prostujących fałszywe informacje, uruchomienie kampanii informacyjnej, czasowe ograniczenie dostępności szkodliwych treści w porozumieniu z platformami cyfrowymi, a także działania legislacyjne i dyplomatyczne w przypadku kampanii dezinformacyjnych sterowanych z zewnątrz. Reagowanie może przybierać formę zarówno operacyjną (np. w ramach zarządzania kryzysowego), jak i długofalową – w postaci reform strukturalnych i zmian w polityce państwa. Kluczowe znaczenie ma tu zdolność do szybkiej koordynacji międzyresortowej oraz opracowanie z góry przygotowanych scenariuszy i procedur działania.

Czwartym, często niedocenianym, ale fundamentalnym filarem systemu jest **edukacja**. To ona kształtuje długoterminową odporność społeczeństwa na zagrożenia informacyjne. Edukacja medialna i cyfrowa powinna być obecna nie tylko w szkołach i na uczelniach, ale także w formie programów społecznych, szkoleń dla urzędników, dziennikarzy, a także w systemie edukacji dorosłych. Jej celem jest budowanie świadomego i krytycznego odbiorcy informacji, zdolnego do rozpoznawania manipulacji, oceny wiarygodności źródeł i świadomego uczestnictwa w debacie publicznej. Edukacja wzmacnia kompetencje indywidualne, które są niezbędne w warunkach demokratycznego społeczeństwa narażonego na szum informacyjny, chaos komunikacyjny i zorganizowane kampanie dezinformacyjne.

System bezpieczeństwa informacyjnego funkcjonuje na **wielu poziomach**, które wzajemnie się przenikają. Na poziomie **narodowym** za jego funkcjonowanie odpowiadają instytucje rządowe i państwowe: Rada Ministrów, odpowiednie ministerstwa (MSWiA, MON, MC), służby specjalne, wojsko oraz organy doradcze i analityczne. Działania te są koordynowane poprzez dokumenty strategiczne, centra operacyjne i mechanizmy wymiany informacji. Na poziomie **sektorowym** bezpieczeństwo informacyjne musi być realizowane wewnątrz instytucji i organizacji – publicznych i prywatnych – zwłaszcza tych odpowiedzialnych za infrastrukturę krytyczną, media,

edukację, zdrowie, finanse czy energetykę. Poziom **lokalny** obejmuje jednostki samorządu terytorialnego, szkoły, biblioteki, lokalne media oraz inicjatywy obywatelskie, które odgrywają istotną rolę w edukacji i reagowaniu na dezinformację w skali mikrospołecznej. Natomiast poziom **międzynarodowy** obejmuje współpracę z organizacjami takimi jak NATO, Unia Europejska, ONZ, a także uczestnictwo w międzynarodowych sieciach wymiany informacji, standardów i najlepszych praktyk.

System bezpieczeństwa informacyjnego nie funkcjonuje w oderwaniu od innych systemów państwowych – przeciwnie, **tworzy z nimi sieć interfejsów** funkcjonalnych i instytucjonalnych. Przede wszystkim ściśle powiązany jest z **systemem bezpieczeństwa narodowego**, będąc jego nieodłączną częścią, szczególnie w obszarze przeciwdziałania zagrożeniom hybrydowym i destabilizacji wewnętrznej. Po drugie, system ten współdziała z mechanizmami **zarządzania kryzysowego**, ponieważ wiele incydentów informacyjnych – zwłaszcza w sytuacjach nadzwyczajnych – wymaga szybkiego reagowania, koordynacji komunikatów i zapewnienia spójności przekazu. Wreszcie, łączy się on z systemem **obrony cywilnej**, ponieważ ochrona informacji i informowanie społeczeństwa są istotnymi elementami ochrony ludności w czasie pokoju, kryzysu i wojny.

Zintegrowana struktura systemu bezpieczeństwa informacyjnego wymaga zarówno dobrze zaprojektowanych rozwiązań organizacyjnych, jak i zdolności do szybkiego adaptowania się do zmieniającego się środowiska zagrożeń. Konieczne jest nie tylko współdziałanie instytucji, ale także jasne przypisanie kompetencji, koordynacja działań, wymiana informacji i wspólne kształtowanie odporności państwa oraz jego obywateli.

2.3. Kluczowe podmioty systemu bezpieczeństwa informacyjnego

System bezpieczeństwa informacyjnego państwa funkcjonuje w oparciu o współdziałanie wielu podmiotów – zarówno instytucji publicznych, jak i aktorów spoza sektora administracyjnego. Każdy z nich pełni odrębną, ale komplementarną funkcję w zakresie monitorowania, przeciwdziałania i reagowania na zagrożenia informacyjne, w tym te wynikające z manipulacji audiowizualnej. W szczególności, w obliczu rozwoju technologii generatywnej, sztucznej inteligencji oraz narzędzi służących do tworzenia

fałszywych treści wideo i audio (np. deepfake), konieczne jest zrozumienie, które instytucje ponoszą odpowiedzialność za ochronę informacyjną państwa i społeczeństwa.

Na poziomie centralnym najważniejsze zadania w zakresie bezpieczeństwa informacyjnego spoczywają na organach władzy wykonawczej, w tym przede wszystkim na Radzie Ministrów i Prezesie Rady Ministrów, który odpowiada za koordynację działań związanych z bezpieczeństwem narodowym, także w wymiarze informacyjnym. Szczególną rolę pełnią tu ministerstwa o charakterze bezpieczeństwa i strategicznym znaczeniu: Ministerstwo Spraw Wewnętrznych i Administracji, Ministerstwo Obrony Narodowej, Ministerstwo Cyfryzacji oraz Ministerstwo Spraw Zagranicznych. W zależności od rodzaju zagrożenia informacyjnego, działania są koordynowane między tymi resortami w ramach mechanizmów zarządzania kryzysowego lub komunikacji strategicznej.

Istotne kompetencje w zakresie wykrywania i przeciwdziałania operacjom dezinformacyjnym, w tym treściom audiowizualnym wykorzystywanym do manipulacji percepcyjnej, posiadają służby specjalne, w szczególności Agencja Bezpieczeństwa Wewnętrznego (ABW), Służba Kontrwywiadu Wojskowego (SKW) oraz Służba Wywiadu Wojskowego (SWW). Zajmują się one zarówno rozpoznawaniem kampanii informacyjnych o charakterze hybrydowym, jak i analizą wpływu obcych służb wywiadowczych na debatę publiczną i opinię społeczną. W ramach ich działalności rośnie znaczenie tzw. kompetencji OSINT (open-source intelligence)[34], czyli pozyskiwania i analizowania informacji z otwartych źródeł, w tym mediów społecznościowych i platform wideo.

Agencja Bezpieczeństwa Wewnętrznego (ABW) jest jednym z kluczowych podmiotów odpowiedzialnych za ochronę bezpieczeństwa narodowego Rzeczypospolitej Polskiej, a jej kompetencje obejmują także istotne obszary bezpieczeństwa informacyjnego. W obliczu rosnącej roli przestrzeni cyfrowej i audiowizualnej jako areny zagrożeń hybrydowych, dezinformacji oraz operacji wpływu, rola ABW w systemie ochrony informacyjnej staje się coraz bardziej istotna i wielowymiarowa.

Podstawowe zadania ABW w odniesieniu do bezpieczeństwa informacyjnego określone są w ustawie o Agencji Bezpieczeństwa Wewnętrznego

[34] M. Witczak, Biały wywiad w zarządzaniu bezpieczeństwem informacji, Przegląd Bezpieczeństwa Wewnętrznego, 30, 214.

oraz Agencji Wywiadu z dnia 24 maja 2002 roku. ABW odpowiada m.in. za: rozpoznawanie, zapobieganie i wykrywanie zagrożeń dla wewnętrznego bezpieczeństwa państwa, ochronę informacji niejawnych, przeciwdziałanie terroryzmowi, ochronę bezpieczeństwa systemów teleinformatycznych oraz przeciwdziałanie zagrożeniom w cyberprzestrzeni[35]. W praktyce oznacza to, że Agencja realizuje działania zarówno w obszarze ochrony technicznej informacji (w tym przeciwdziałania atakom cybernetycznym), jak i ochrony przestrzeni komunikacyjnej przed manipulacją i wpływami zewnętrznymi.

W zakresie przeciwdziałania zagrożeniom informacyjnym ABW prowadzi działania analityczne i operacyjne mające na celu rozpoznawanie i neutralizację kampanii dezinformacyjnych, których celem jest destabilizacja wewnętrzna państwa, polaryzacja społeczna, podważanie zaufania do instytucji publicznych lub ingerencja w procesy demokratyczne, w tym wybory. Działania te obejmują także analizę treści audiowizualnych – takich jak zmanipulowane nagrania wideo, deepfake'i, sfałszowane przekazy medialne – które są wykorzystywane do budowania fałszywych narracji i oddziaływania na emocje społeczne.

Jednym z ważnych obszarów działalności ABW jest ochrona państwowej przestrzeni informacyjnej przed operacjami wpływu prowadzonymi przez podmioty zagraniczne, w tym obce służby specjalne. Agencja monitoruje aktywność podmiotów medialnych, organizacji społecznych i innych struktur mogących stanowić instrumenty wpływu zewnętrznego, a także przeciwdziała kampaniom dezinformacyjnym mającym na celu podważenie spójności społecznej, rozpowszechnianie treści propagandowych lub destabilizację działań rządu i instytucji demokratycznych.

W ramach ochrony systemów teleinformatycznych i infrastruktury krytycznej ABW prowadzi działania prewencyjne i reagujące na incydenty bezpieczeństwa dotyczące sieci rządowych, systemów administracji publicznej, a także sektorów strategicznych, takich jak energetyka, zdrowie, transport czy finanse. Ochrona techniczna informacji obejmuje zarówno przeciwdziałanie bezpośrednim cyberatakom, jak i zapobieganie wyciekom danych, które mogłyby zostać wykorzystane w kampaniach dezinformacyjnych.

Agencja odgrywa także istotną rolę w procesie współpracy międzynarodowej w zakresie bezpieczeństwa informacyjnego i cyberbezpieczeństwa,

[35] T. M. Miłkowski, Czy ABW powinna korzystać ze „specjalnego" prawa do informacji, Przegląd Bezpieczeństwa Wewnętrznego, 10, 211.

uczestnicząc w wymianie informacji, analiz i dobrych praktyk w ramach struktur NATO, Unii Europejskiej oraz bilateralnych porozumień z innymi państwami partnerskimi. Prowadzenie współpracy międzynarodowej w obszarze wymiany informacji niejawnych wymaga stworzenia podstaw formalnoprawnych, które umożliwią określenie zasad, na podstawie których prowadzona będzie wymiana informacji niejawnych. Instrumentem prawnym służącym prowadzeniu takiej współpracy przez Rząd Rzeczypospolitej Polskiej ze swoimi zagranicznymi odpowiednikami są umowy międzynarodowe. Należy również wskazać, że właściwość Szefa ABW w zakresie czynności związanych z zawieraniem umów międzynarodowych, dotyczących wzajemnej ochrony i wymiany informacji niejawnych, wynika z następujących przepisów:

- art. 5 ustawy z dnia 14 kwietnia 2000 r. o umowach międzynarodowych (Dz.U. 2020, poz. 127 z późn. zm.) — minister kierujący działem administracji rządowej właściwy do spraw, których dotyczy umowa międzynarodowa, po uzgodnieniu projektu umowy i instrukcji negocjacyjnej z ministrem właściwym do spraw zagranicznych oraz z innymi zainteresowanymi ministrami, składa Prezesowi Rady Ministrów wniosek o udzielenie zgody na rozpoczęcie negocjacji umowy międzynarodowej. Przepisy ustawy odnoszące się do ministra stosuje się odpowiednio do kierownika centralnego organu administracji rządowej
- (art. 4 ustawy);
- art. 5 ust. 1 pkt 3 ustawy o ABW oraz AW — do zadań ABW należy realizowanie, w granicach swojej właściwości, zadań związanych z ochroną informacji niejawnych oraz wykonywanie funkcji krajowej władzy bezpieczeństwa w zakresie ochrony informacji niejawnych w stosunkach międzynarodowych;
- art. 11 ust. u. inf. niejaw. — Szef ABW pełni funkcję krajowej władzy bezpieczeństwa[36].

[36] M. Nowiński, M. Wieczerza, Zadania służb specjalnych w obszarze ochrony informacji niejawnych ze szczególnym uwzględnieniem aspektów międzynarodowych [w:] Prawne aspekty funkcjonowania służb specjalnych na przykładzie Agencji Bezpieczeństwa Wewnętrznego, P. Burczaniuk (red.), Wydawnictwo Agencji Bezpieczeństwa Wewnętrznego, Warszawa 2021, s. 131.

Dzięki temu ABW uzyskuje dostęp do bieżących informacji o nowych zagrożeniach informacyjnych, technikach manipulacji treściami audiowizualnymi oraz strategiach przeciwdziałania kampaniom destabilizacyjnym.

Istotnym aspektem działalności ABW w obszarze bezpieczeństwa informacyjnego jest również współpraca z mediami publicznymi i sektorem prywatnym. Agencja współpracuje z operatorami platform cyfrowych, dostawcami usług internetowych oraz podmiotami odpowiedzialnymi za infrastrukturę informacyjną w zakresie identyfikacji i neutralizacji zagrożeń, reagowania na incydenty oraz prowadzenia działań informacyjnych i edukacyjnych.

Współczesne wyzwania wymagają od ABW nie tylko zdolności technicznych, ale także rozwijania kompetencji w zakresie analizy narracji audiowizualnych, psychologii komunikacji oraz nowych technologii informacyjnych, takich jak sztuczna inteligencja czy analiza danych wielkoskalowych (big data). W tym kontekście istotne jest wzmacnianie zdolności Agencji do monitorowania i przeciwdziałania zagrożeniom wynikającym z masowego rozpowszechniania zmanipulowanych treści wizualnych, które stają się narzędziem walki informacyjnej na skalę globalną.

Agencja Bezpieczeństwa Wewnętrznego pełni kluczową funkcję w systemie bezpieczeństwa informacyjnego państwa, łącząc działania prewencyjne, analityczne i operacyjne w celu ochrony społeczeństwa i instytucji przed zagrożeniami płynącymi z przestrzeni informacyjnej – zwłaszcza tymi, które wykorzystują treści audiowizualne do manipulacji, dezinformacji i destabilizacji.

Współczesne zagrożenia informacyjne – zwłaszcza te związane z manipulacją audiowizualną, dezinformacją i operacjami wpływu – mają charakter zarówno cywilny, jak i wojskowy. Dlatego też w systemie bezpieczeństwa informacyjnego Rzeczypospolitej Polskiej szczególną rolę odgrywają Służba Kontrwywiadu Wojskowego (SKW) oraz Służba Wywiadu Wojskowego (SWW). Są one odpowiedzialne za rozpoznawanie, przeciwdziałanie i neutralizowanie zagrożeń informacyjnych w sferze wojskowej, strategicznej i międzynarodowej, zwłaszcza w kontekście ochrony Sił Zbrojnych RP, struktur sojuszniczych oraz operacji prowadzonych poza granicami kraju.

Służba Kontrwywiadu Wojskowego (SKW) pełni kluczową rolę w ochronie wewnętrznej bezpieczeństwa informacyjnego wojska oraz infrastruktury obronnej państwa. Jej zadania obejmują wykrywanie

i przeciwdziałanie działalności obcych służb wywiadowczych, także w przestrzeni informacyjnej. W kontekście zagrożeń audiowizualnych SKW analizuje m.in. działania mające na celu dezinformowanie lub kompromitowanie Sił Zbrojnych RP za pomocą spreparowanych materiałów wideo, deepfake-'ów przedstawiających fałszywe zdarzenia militarne, czy kampanii propagandowych wymierzonych w morale żołnierzy oraz opinię publiczną.

Współczesne operacje kontrwywiadowcze muszą uwzględniać techniki prowadzenia operacji wpływu przy użyciu treści audiowizualnych – takich jak sfałszowane nagrania ukazujące rzekome nieprawidłowości w funkcjonowaniu jednostek wojskowych, dezinformacyjne filmy o działaniach polskich żołnierzy za granicą, czy zmanipulowane relacje mające na celu podważenie zaufania społeczeństwa do instytucji obronnych. W tym kontekście SKW realizuje działania prewencyjne, analityczne i przeciwdziałające, chroniąc jednocześnie sieci komunikacyjne i informacyjne wykorzystywane w ramach systemu dowodzenia i łączności sił zbrojnych.

Służba Wywiadu Wojskowego (SWW) odpowiada natomiast za rozpoznawanie zagrożeń informacyjnych poza granicami Polski, w tym działalności wywiadowczej, dezinformacyjnej oraz propagandowej, które mogą mieć wpływ na bezpieczeństwo państwa. SWW zbiera informacje o planowanych lub realizowanych operacjach informacyjnych państw trzecich, analizuje trendy w wykorzystaniu treści audiowizualnych do destabilizacji sytuacji międzynarodowej, a także ostrzega władze RP o potencjalnych kampaniach manipulacyjnych wymierzonych w polskie interesy.

W szczególności SWW monitoruje wykorzystanie nowoczesnych narzędzi – takich jak generatywne modele sztucznej inteligencji tworzące fałszywe materiały filmowe – przez potencjalnych przeciwników, zarówno w ramach operacji hybrydowych, jak i w kontekście wojen psychologicznych. Agencja ta odgrywa istotną rolę w identyfikowaniu kampanii dezinformacyjnych prowadzonych na obszarach działań polskich kontyngentów wojskowych oraz w przestrzeni sojuszniczej (NATO, UE).

Zarówno SKW, jak i SWW uczestniczą w budowaniu narodowego systemu odporności informacyjnej poprzez wymianę informacji z innymi służbami specjalnymi, udział w międzynarodowych strukturach bezpieczeństwa (np. NATO Intelligence Fusion Centre), a także poprzez dostarczanie analiz zagrożeń dla sztabów wojskowych, struktur operacyjnych i decydentów politycznych. Współpraca obu służb obejmuje również działania

w obszarze ochrony informacji niejawnych i systemów teleinformatycznych wykorzystywanych w obronności. W czasach, gdy kampanie dezinformacyjne często opierają się na wycieku lub manipulacji wizualnej danych (np. sfałszowane dokumenty, materiały wideo rzekomo pochodzące z operacji wojskowych), rola SKW i SWW w zabezpieczaniu integralności informacyjnej państwa jest niezwykle istotna. W perspektywie strategicznej działalność SKW i SWW w systemie bezpieczeństwa informacyjnego koncentruje się na:

- wczesnym rozpoznawaniu zagrożeń narracyjnych i operacji wpływu,
- ochronie morale i wizerunku Sił Zbrojnych RP w przestrzeni medialnej,
- analizie i przeciwdziałaniu zagranicznym operacjom dezinformacyjnym prowadzonym przy użyciu treści audiowizualnych,
- zapewnieniu bezpieczeństwa sieci teleinformatycznych i systemów komunikacji strategicznej,
- wspieraniu struktur NATO i UE w budowie wspólnej odporności informacyjnej.

Służba Kontrwywiadu Wojskowego i Służba Wywiadu Wojskowego stanowią integralny filar systemu bezpieczeństwa informacyjnego państwa, zwłaszcza w kontekście ochrony wojskowej przestrzeni komunikacyjnej, przeciwdziałania manipulacji audiowizualnej oraz neutralizowania zagrożeń hybrydowych na poziomie operacyjnym i strategicznym. Ich działalność – prowadzona zarówno w kraju, jak i za granicą – ma kluczowe znaczenie dla utrzymania suwerenności informacyjnej i skuteczności działań obronnych Polski[37].

W obszarze monitorowania i technicznej ochrony zasobów informacyjnych funkcjonuje CERT Polska (zespół reagowania na incydenty komputerowe) działający w strukturach NASK – państwowego instytutu badawczego zajmującego się również analizą dezinformacji i rozwojem narzędzi do jej wykrywania. CERT-y funkcjonują również przy innych instytucjach publicznych oraz w sektorze prywatnym, zapewniając wsparcie w zakresie cyberbezpieczeństwa, które bezpośrednio wpływa na bezpieczeństwo informacyjne.

[37] R. Brzozowski, Czynności wykonywane przez funkcjonariuszy ABW na tle zadań ABW [w:] Prawne aspekty funkcjonowania służb specjalnych na przykładzie Agencji Bezpieczeństwa Wewnętrznego, P. Burczaniuk (red.), Wydawnictwo Agencji Bezpieczeństwa Wewnętrznego, Warszawa 2021, s. 147.

Ważną rolę pełni także Rządowe Centrum Bezpieczeństwa (RCB), które odpowiada za koordynację działań w sytuacjach kryzysowych, w tym za informowanie społeczeństwa o zagrożeniach i przeciwdziałanie panice informacyjnej[38]. RCB jako instytucja o charakterze komunikacyjnym, współpracuje z innymi organami w celu zapewnienia spójnego, wiarygodnego i szybkiego przekazu – co jest niezwykle istotne zwłaszcza w sytuacjach, gdy dezinformacja rozprzestrzenia się w formie materiałów audiowizualnych.

W systemie bezpieczeństwa informacyjnego istotne miejsce zajmują również media publiczne, takie jak Telewizja Polska, Polskie Radio i Polska Agencja Prasowa, które zobowiązane są do pełnienia misji publicznej. Oznacza to nie tylko przekazywanie rzetelnej informacji, ale także reagowanie na fałszywe narracje, prostowanie nieprawdziwych informacji i wspieranie edukacji obywatelskiej. Media publiczne powinny także uczestniczyć w systemie wczesnego ostrzegania o zagrożeniach informacyjnych oraz współpracować z instytucjami państwowymi w sytuacjach kryzysowych.

Nie sposób pominąć roli sektora prywatnego, zwłaszcza platform cyfrowych (takich jak Meta, YouTube, TikTok czy X), które zarządzają głównymi kanałami dystrybucji treści audiowizualnych. Ich działania – bądź zaniechania – mają bezpośredni wpływ na to, jakie informacje trafiają do odbiorców, w jakiej formie i z jaką siłą oddziaływania. Od platform oczekuje się współpracy z organami państwowymi, transparentności algorytmów rekomendujących treści oraz szybkiego reagowania na zgłoszenia dotyczące treści manipulacyjnych lub fałszywych. Równocześnie coraz częściej pojawiają się postulaty prawnego uregulowania ich odpowiedzialności za bezpieczeństwo informacyjne użytkowników.

Istotnym aktorem w systemie bezpieczeństwa informacyjnego są również organizacje pozarządowe, ośrodki eksperckie i środowiska akademickie. Pełnią one rolę watchdogów, prowadzą niezależną analizę przekazów medialnych, tworzą raporty o zagrożeniach narracyjnych i dezinformacyjnych, rozwijają narzędzia fact-checkingowe oraz uczestniczą w kampaniach edukacyjnych. Instytucje te, takie jak np. Fundacja Demagog, Stowarzyszenie 61, NASK PIB czy Centrum Badań nad Terroryzmem Collegium Civitas,

[38] W. Skomra, Miejsce, rola i zadania Rządowego Centrum Bezpieczeństwa w zakresie integracji działań na rzecz utrzymania bezpieczeństwa [w:] Ratownictwo Wodne oraz inne aspekty bezpieczeństwa na obszarach wodnych. Dylematy i wyzwania, J. Telak (red.), Centrum Szkolenia Policji, Legionowo 2017, s. 16.

wspierają państwo w wykrywaniu i opisie nowych form manipulacji audiowizualnej oraz służą rekomendacjami w zakresie przeciwdziałania.

Na koniec należy podkreślić, że skuteczność systemu bezpieczeństwa informacyjnego zależy również od zaangażowania społeczeństwa obywatelskiego – świadomego, odpornego na manipulację i aktywnego w przeciwdziałaniu dezinformacji. Obywatele mogą odgrywać kluczową rolę w identyfikowaniu i zgłaszaniu treści wprowadzających w błąd, wspieraniu niezależnych mediów oraz uczestnictwie w działaniach edukacyjnych i społecznych.

Podsumowując, system bezpieczeństwa informacyjnego opiera się na współdziałaniu wielu podmiotów, których kompetencje i możliwości muszą być jasno zdefiniowane, dobrze skoordynowane i stale rozwijane. W obliczu szybko rosnącej roli treści audiowizualnych – jako nośników dezinformacji, propagandy i operacji wpływu – współpraca między instytucjami państwowymi, mediami, sektorem technologicznym i obywatelami staje się warunkiem koniecznym dla ochrony informacyjnej suwerenności państwa.

2.4. Zagrożenia dla bezpieczeństwa informacyjnego

Współczesne bezpieczeństwo informacyjne państwa znajduje się pod stałą presją dynamicznie zmieniających się zagrożeń, które mają charakter zarówno wewnętrzny, jak i zewnętrzny, cyfrowy i kulturowy, technologiczny i narracyjny. Szczególną rolę we współczesnej architekturze zagrożeń odgrywają treści audiowizualne, które – dzięki swojej atrakcyjności, sile emocjonalnego przekazu oraz możliwościom manipulacji – stały się podstawowym nośnikiem działań dezinformacyjnych i destabilizujących. Odbiorca obrazu i dźwięku jest zazwyczaj mniej krytyczny niż wobec treści tekstowych, co czyni przekazy audiowizualne skutecznym narzędziem wpływu psychologicznego i socjotechnicznego. Za warunki osiągnięcia bezpieczeństwa informacyjnego w literaturze przedmiotu przyjmuję się:

- niezagrożone strategiczne zasoby państwa;
- decyzje organów władzy podjęte na podstawie wiarygodnych, istotnych informacji;
- niezakłócony przepływ informacji pomiędzy organami państwa;

- niezakłócone funkcjonowanie sieci teleinformatycznych tworzących krytyczną infrastrukturę teleinformatyczną państwa;
- zagwarantowaną przez państwo ochronę informacji niejawnych
- i danych osobowych obywateli;
- zasadę, że prawo do prywatności obywateli jest nienaruszane przez instytucje publiczne,
- swobodny dostęp obywateli do informacji publicznej[39].

Jednym z najbardziej powszechnych i jednocześnie niebezpiecznych zjawisk pozostaje dezinformacja, czyli celowe rozpowszechnianie fałszywych lub zmanipulowanych informacji w celu wprowadzenia odbiorców w błąd. W kontekście audiowizualnym dezinformacja przybiera formę zmanipulowanych filmów, fałszywych materiałów wideo „z miejsca wydarzenia", animacji i grafik, które imitują relacje medialne lub wypowiedzi osób publicznych. Często są one produkowane przez wyspecjalizowane ośrodki propagandowe lub zautomatyzowane systemy wykorzystujące sztuczną inteligencję, a ich celem jest wywołanie paniki, zniekształcenie obrazu rzeczywistości lub podważenie zaufania do instytucji państwowych i mediów.

Kolejną kategorią zagrożeń są operacje wpływu (influence operations), prowadzone przez aktorów państwowych lub niepaństwowych, których celem jest strategiczne oddziaływanie na percepcję społeczną, emocje zbiorowe oraz decyzje polityczne[40]. Operacje te bardzo często wykorzystują treści audiowizualne – w tym spoty, inscenizacje, pseudodokumenty i sfałszowane nagrania – jako narzędzie kształtowania narracji korzystnych dla interesów nadawcy. Działania tego typu mają charakter długofalowy, są trudne do jednoznacznego przypisania, a ich skutki bywają odczuwalne dopiero po czasie. Wzrost znaczenia tzw. „miękkich" form wojny informacyjnej powoduje, że analiza treści wizualnych staje się kluczowym elementem strategii bezpieczeństwa narodowego.

Szczególnie niepokojącym zjawiskiem w ostatnich latach jest rozwój technologii deepfake, umożliwiającej tworzenie realistycznych, lecz całkowicie fałszywych nagrań wideo i audio, przedstawiających np. polityków,

[39] E. Nowak, M. Nowak, Zarys teorii bezpieczeństwa narodowego, Difin SA, Warszawa 2011, s. 103.

[40] P. Brangetto, M. A. Veenendaal, Influence Cyber Operations: The use of cyberattacks in support of Influence Operations,2016 8th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2016, s. 115.

dowódców wojskowych czy znane postaci publiczne wypowiadających słowa, których nigdy nie wypowiedzieli. Tego typu treści są nie tylko trudne do wykrycia przez przeciętnego odbiorcę, ale również potencjalnie destabilizujące – mogą wpłynąć na wynik wyborów, wywołać zamieszki społeczne lub sprowokować działania militarne. W połączeniu z szybką dystrybucją w mediach społecznościowych, deepfake'i stają się narzędziem o sile rażenia porównywalnej z bronią psychologiczną.

Do zagrożeń informacyjnych należy zaliczyć również zjawiska o charakterze techniczno-operacyjnym, takie jak phishing i visual phishing, czyli ataki wykorzystujące zmanipulowane obrazy lub interfejsy aplikacji w celu wyłudzenia danych, zainfekowania urządzeń lub zmylenia użytkownika. Ataki te często imitują komunikaty instytucji publicznych lub mediów i rozprzestrzeniają się za pośrednictwem materiałów wideo, reklam lub sfałszowanych transmisji na żywo.

W obszarze społecznym nie można pominąć zjawiska polaryzacji i radykalizacji informacyjnej, napędzanej przez algorytmy platform społecznościowych, które promują treści kontrowersyjne, emocjonalne i często skrajne. Treści audiowizualne – zwłaszcza krótkie formy wideo i memy – odgrywają w tym procesie centralną rolę, ponieważ są łatwo przyswajalne, wywołują szybkie reakcje i łatwo się rozprzestrzeniają. W rezultacie przestrzeń informacyjna ulega fragmentaryzacji, a społeczeństwo staje się podatne na konflikty narracyjne, manipulacje tożsamościowe i dezintegrację komunikacyjną.

Kontekst międzynarodowy wskazuje na coraz częstsze wykorzystywanie narracji strategicznych w konfliktach politycznych i militarnych. Państwa takie jak Rosja, Chiny czy Iran wykorzystują treści wizualne jako komponent wojny informacyjnej – publikując spreparowane relacje, zafałszowane obrazy z pól walki, emocjonalne materiały przedstawiające „ofiary agresji" lub wizerunki „bohaterów narodowych". Celem takich działań jest z jednej strony wpłynięcie na opinię publiczną we własnym kraju, z drugiej – zaburzenie postrzegania sytuacji przez społeczność międzynarodową. W warunkach wojny hybrydowej, obraz staje się bronią – używaną do osłabienia morale przeciwnika, dezinformowania opinii publicznej i prowadzenia działań psychologicznych na masową skalę.

Niebezpieczeństwo dla bezpieczeństwa informacyjnego wynika także z niskiej świadomości społecznej, deficytów edukacji medialnej oraz braku

umiejętności weryfikowania źródeł. Obywatele, niezdolni do rozpoznania fałszu, często nieświadomie stają się przekaźnikami dezinformacji, zwłaszcza gdy materiały są atrakcyjne wizualnie i odpowiadają ich przekonaniom. W tym sensie słabość odporności społecznej jest jednym z najpoważniejszych zagrożeń – nie tylko jako efekt manipulacji zewnętrznej, ale jako systemowe wyzwanie dla funkcjonowania demokracji i instytucji publicznych.

System bezpieczeństwa informacyjnego państwa musi uwzględniać złożony katalog, bez wątpienia należy uwzględnić, iż pojęcie bezpieczeństwa informacyjnego stosuje się także do informacji spoza systemu teleinformatycznego zagrożeń[41], lecz coraz częściej wspólnym mianownikiem stają się coraz częściej treści audiowizualne. Ich siła perswazyjna, zasięg i podatność na manipulację czynią z nich potężne narzędzie oddziaływania – zarówno w rękach państw wrogich, jak i nieodpowiedzialnych aktorów wewnętrznych. Zrozumienie tych zagrożeń i odpowiednie dostosowanie strategii reagowania stanowi podstawę skutecznego systemu ochrony informacyjnej państwa w XXI wieku.

[41] L. Więcaszek-Kuczyńska, Zagrożenia bezpieczeństwa informacyjnego, Obronność – Zeszyty Naukowe Wydziału Zarządzania i Dowodzenia Akademii Sztuki Wojennej, 2(10), 216.

III

Rola analizy treści audiowizualnych w systemach bezpieczeństwa państwa

3.1. Znaczenie analizy treści audiowizualnych

Analiza treści audiowizualnych oznacza systematyczne, empiryczne i możliwie obiektywne badanie przekazu zawartego w obrazach, nagraniach audio oraz materiałach wideo. W literaturze naukowej podkreśla się, że podstaw tego rodzaju badań należy szukać w klasycznych metodach analizy zawartości mediów (ang. *content analysis*), które koncentrują się na identyfikacji wzorców, znaczeń oraz kontekstów występujących w przekazie medialnym[42]. O ile jednak tradycyjna analiza treści skupiała się głównie na materiale tekstowym (np. artykułach prasowych, wypowiedziach polityków czy dokumentach historycznych), to rozwój technologii cyfrowych i popularyzacja mediów elektronicznych stworzyły potrzebę rozwoju narzędzi zdolnych do przetwarzania wielowymiarowych danych audiowizualnych[43]. W efekcie współczesna analiza treści audiowizualnych łączy metody zaczerpnięte z nauk humanistycznych (medioznawstwo, kulturoznawstwo, filmoznawstwo) z podejściami charakterystycznymi dla nauk ścisłych (informatyka, matematyka, inżynieria danych), tworząc interdyscyplinarną dziedzinę badań[44].

[42] Zob. K. Krippendorff, Content Analysis: An Introduction to Its Methodology, 4th ed., SAGE Publications, Thousand Oaks 2018.

[43] Zob. L. A. Neuendorf, The Content Analysis Guidebook, 2nd ed., SAGE Publications, Thousand Oaks 2016.

[44] Zob. B. Dobek-Ostrowska, Podstawy komunikowania społecznego, Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław 2004.

W kontekście definicji analizy treści audiowizualnych warto zaznaczyć, że chodzi nie tylko o rejestrowanie tego, co jest prezentowane w danym obrazie czy nagraniu, ale również w jaki sposób przekaz został skonstruowany, zinterpretowany i odebrany przez odbiorców[45]. Procedura badawcza zwykle obejmuje kilka etapów:

- Selekcja materiału (np. wybór określonej próbki nagrań wideo, serii zdjęć czy nagrań audio),
- Operacjonalizacja kategorii analitycznych (ustalenie, jakie cechy formalne lub treściowe będą poddawane kodowaniu),
- Kodowanie materiału (przez zespół badaczy lub z wykorzystaniem narzędzi automatycznych, np. algorytmów rozpoznawania obrazu i dźwięku),
- Interpretacja wyników (weryfikacja hipotez badawczych, identyfikacja wzorców, wnioskowanie na temat potencjalnego wpływu na odbiorców).

Takie wieloetapowe podejście pozwala na uchwycenie zarówno warstwy jawnej (co zostało pokazane lub powiedziane), jak i ukrytej (symbolika, sugestie, techniki manipulacji) przekazu audiowizualnego[46].

Współczesne bezpieczeństwo państwa coraz częściej definiowane jest w kategoriach zdolności do ochrony nie tylko terytorium i obywateli, lecz także przestrzeni informacyjnej, w której obywatele funkcjonują[47]. Rozwój technologii cyfrowych sprawił, że informacje w postaci audiowizualnej – filmy, materiały dźwiękowe, nagrania transmisji na żywo – rozprzestrzeniają się w tempie trudnym do zarejestrowania i kontrolowania przez tradycyjne instytucje państwowe[48]. Szacuje się, że na samym YouTube co minutę pojawia się około 500 godzin nowych materiałów wideo, co daje ponad 30 tysięcy godzin nagrań w ciągu godziny. Dodatkowo, na platformach społecznościowych dziennie publikowane są miliony materiałów fotograficznych,

[45] Zob. M. B. Saldaña, The Coding Manual for Qualitative Researchers, 3rd ed., SAGE Publications, Los Angeles 2016.

[46] Zob. T. Goban-Klas, Media i komunikowanie masowe, Wydawnictwo Naukowe PWN, Warszawa 2009, s. 145-162.

[47] Zob. J. Nye, Soft Power: The Means to Success in World Politics, PublicAffairs, New York 2004.

[48] Zob. Y. Li, M.C. Chang, S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong 2018, s. 1-7.

filmowych i podcastów, z czego wiele może zawierać treści potencjalnie niebezpieczne dla stabilności państwa (np. nawołujące do przemocy, szerzące nienawiść rasową czy religijną, dezinformujące opinię publiczną).

Analiza treści audiowizualnych nabiera w związku z tym szczególnego znaczenia w dwóch głównych obszarach. Pierwszym z nich jest wczesne ostrzeganie i prewencja[49]. Służby państwowe, takie jak agencje wywiadowcze czy organy ścigania, wykorzystują zaawansowane narzędzia do monitorowania ogólnodostępnych przekazów audiowizualnych w celu wykrycia sygnałów świadczących o narastających zagrożeniach informacyjnych, takich jak agitacja terrorystyczna, komunikacja ekstremistyczna czy kampanie dezinformacyjne prowadzone przez obce państwa[50]. Na przykład organizacje terrorystyczne regularnie posługują się materiałami wideo służącymi propagandzie, rekrutacji nowych członków oraz wywieraniu presji psychologicznej na przeciwników. Systematyczna analiza takich treści, obejmująca rozpoznawanie wzorców językowych, symboliki oraz używanych środków artystycznych, pozwala na wcześniejszą identyfikację podmiotów zaangażowanych w działalność nielegalną oraz antycypację możliwych ataków terrorystycznych[51]. Drugim kluczowym obszarem jest diagnoza wpływu na społeczeństwo i stabilność państwa. Audiowizualne formy komunikacji, poprzez warstwę wizualną i dźwiękową, silnie oddziałują na emocje odbiorców, przyciągając uwagę znacznie skuteczniej niż tradycyjny tekst[52]. Materiały te mogą kształtować opinie publiczne oraz postawy polityczne, szczególnie gdy są szeroko rozpowszechniane w mediach społecznościowych. W epoce tzw. post-prawdy, gdzie granica między informacją a dezinformacją ulega zatarciu, umiejętność szybkiego odróżnienia materiałów autentycznych od zmanipulowanych ma zasadnicze znaczenie. Szczególnie istotne jest to w sytuacjach kryzysowych, takich jak konflikty zbrojne, ataki terrorystyczne czy klęski żywiołowe, gdzie fałszywe nagrania czy spreparowane zdjęcia

[49] Rezolucja Parlamentu Europejskiego z dnia 9 marca 2022 r. w sprawie obcych ingerencji we wszystkie procesy demokratyczne w Unii Europejskiej, w tym dezinformacji (2020/2268(INI)), Dz.U. UE C 347 z 9.9.2022, s. 61.

[50] Zob. N. Jin, Big Data Analysis of National Security based on Network Traffic Metadata, IEEE, 2023.

[51] Zob. T. Bąk, Rola mediów we współczesnych konfliktach zbrojnych i aktach terrorystycznych, WSIiZ, 2020.

[52] Zob. T. Goban-Klas, Media i komunikowanie masowe, Wydawnictwo Naukowe PWN, Warszawa 2009, s. 145-162.

mogą wywołać panikę, pogłębiać niepokoje społeczne lub podsycać konflikty między grupami etnicznymi czy religijnymi.

Na znaczenie analizy treści audiowizualnych wpływa również fakt, że współczesne kampanie dezinformacyjne coraz częściej przyjmują formę zaawansowanych technicznie manipulacji określanych jako *deepfake*, które wykorzystują algorytmy uczenia głębokiego do generowania fałszywych nagrań audio lub wideo[53]. Dzięki tym technologiom możliwe jest tworzenie realistycznych, choć fałszywych wypowiedzi polityków, wojskowych czy osób publicznych, co może powodować dezorientację społeczną. Identyfikacja i weryfikacja autentyczności takich materiałów wymaga zastosowania zaawansowanych metod analizy, takich jak wyspecjalizowane sieci neuronowe służące do wykrywania subtelnych anomalii niedostrzegalnych ludzkim okiem. W Strategii Bezpieczeństwa Narodowego RP z 2020 roku wyraźnie podkreślono konieczność wzmacniania zdolności państwa do przeciwdziałania dezinformacji i propagandzie, w tym rozwijania narzędzi służących do analizy i filtrowania treści audiowizualnych[54].

Warto zaznaczyć, że analiza treści audiowizualnych wykracza poza kwestie bezpieczeństwa, stanowiąc istotny element badań naukowych nad komunikacją społeczną, kulturą wizualną oraz wpływem mediów na procesy polityczne i gospodarcze. Jednak w kontekście bezpieczeństwa państwa jej rola jest szczególnie ważna – od identyfikacji potencjalnych zagrożeń i wspierania procesów decyzyjnych w sytuacjach kryzysowych, aż po kształtowanie długoterminowej strategii obrony informacyjnej. Kluczowym wyzwaniem pozostaje integracja narzędzi analizy semantycznej, rozpoznawania mowy oraz wizji komputerowej w jednolity system monitoringu[55]. Jednocześnie niezwykle istotne są aspekty etyczne i prawne, takie jak ochrona prywatności obywateli, wolność słowa czy ryzyko nadmiernej inwigilacji społeczeństwa[56]. Podsumowując, analiza treści audiowizualnych to złożony i wielowymiarowy proces, który w erze cyfrowej odgrywa rosnącą rolę w systemach bezpieczeństwa państwa. Jej znaczenie polega głównie na

---

[53] Zob. M. Tabakow, J. Korczak, B. Franczyk, Big Data – definicje, wyzwania i technologie informatyczne, Informatyka Ekonomiczna, 1(31).

[54] Strategia Bezpieczeństwa Narodowego RP, Biuro Bezpieczeństwa Narodowego, 2020.

[55] Zob. J. Kurek, Bezpieczeństwo państwa w warunkach hybrydowej regulacji danych osobowych w dobie analizy Big Data, ASzWoj, Warszawa 2021.

[56] Zob. D. Boyd, K. Crawford, Critical Questions for Big Data, Information, Communication & Society, 15(5), 2012.

możliwości wczesnego wykrywania zagrożeń informacyjnych, diagnozowania wpływu przekazów medialnych na społeczeństwo oraz skutecznego reagowania na pojawiające się wyzwania. Interdyscyplinarność tej analizy wymaga łączenia kompetencji społecznych, humanistycznych oraz technologicznych, co umożliwia coraz bardziej skuteczne identyfikowanie manipulacji i przewidywanie trendów w kształtowaniu opinii publicznej.

3.2. Multimodalność w analizie zagrożeń informacyjnych

Współczesna sfera informacyjna jest zdominowana przez przekazy łączące różnorodne formy ekspresji: tekst, obraz statyczny, nagrania wideo oraz dźwięk. W konsekwencji dezinformacja i propaganda rzadko ograniczają się wyłącznie do słów, dla większej skuteczności i wiarygodności manipulacji często stosuje się elementy audiowizualne, wzbogacone o odpowiednio skonstruowaną narrację tekstową. Dlatego analizowanie jedynie jednej modalności – na przykład wyłącznie warstwy językowej – okazuje się niewystarczające do wykrycia potencjalnych manipulacji i zrozumienia szerszego kontekstu przekazu. W literaturze przedmiotu podkreśla się, że dopiero analiza wielomodalna, uwzględniająca słowo, obraz i dźwięk w ich wzajemnej relacji, pozwala wychwycić zaawansowane strategie perswazyjne stosowane przez podmioty prowadzące operacje dezinformacyjne.

W badaniach nad dezinformacją i propagandą coraz częściej zwraca się uwagę na konieczność uwzględniania różnych form przekazu. W związku z tym, skuteczne przeciwdziałanie dezinformacji wymaga integracji metod analizy tekstu, obrazu i dźwięku, co pozwala na pełniejsze zrozumienie mechanizmów manipulacji i skuteczniejsze wykrywanie fałszywych treści.

Multimodalność stanowi kluczową koncepcję w badaniach nad komunikacją, opierając się na założeniu, że porozumiewanie się ludzi odbywa się w wielu kanałach i systemach semiotycznych równocześnie[57]. Materiały wideo publikowane w Internecie, czy szerzej – przekazy medialne – mogą zawierać zarówno napisy, jak i symbole graficzne czy ścieżkę dźwiękową o wyrazistym charakterze. Analiza zjawisk związanych z bezpieczeństwem informacyjnym wymaga zatem przyjęcia perspektywy, w której warstwa

[57] Zob. G. Kress, T. van Leeuwen, Reading Images: The Grammar of Visual Design, Taylor i Francis, 2020.

tekstu, obrazu i dźwięku jest rozpatrywana łącznie[58]. W operacjach manipulacyjnych często łączy się bowiem różne środki wyrazu, a manipulacja może tkwić w interakcjach między nimi. W ujęciu nauk o bezpieczeństwie państwa szczególnie istotne jest, aby narzędzia analizy informacji wychwytywały nie tylko fałszywe lub zniekształcone opisy zdarzeń, ale też np. niuanse kompozycji kadru czy ładunek emocjonalny ścieżki audio, mogący silnie oddziaływać na odbiorców.

W kontekście wojskowym i cywilnym dezinformacja przybiera często postać skoordynowanych operacji psychologicznych, zwanych też psy-ops, w których łączy się różnorodne formy komunikacji w jednolitą, choć zmanipulowaną, narrację. Aby rozpoznać złożone techniki używane w tego rodzaju kampaniach, analitycy sięgają po narzędzia tzw. *cross-modal analysis*, polegające na badaniu spójności i konsekwencji między tekstem, obrazem a dźwiękiem[59]. Szczególne znaczenie w ostatnich latach zyskało zagadnienie *deepfake*, czyli syntetycznych materiałów audiowizualnych generowanych z wykorzystaniem sieci (ang. *Generative Adversarial Network*, GAN)[60]. Metody te pozwalają na stworzenie wideo prezentującego osoby publiczne wypowiadające zdania, których nigdy nie wypowiedziały[61]. Takie nagrania mogą nie tylko wprowadzać w błąd opinię publiczną, lecz także wywoływać poważne skutki destabilizujące życie polityczne i społeczne. Właśnie w tym obszarze analiza multimodalna wykazuje wyjątkową przydatność. Efektywne metody detekcji *deepfake* łączą tradycyjne podejście wizualne, skupione na wykrywaniu artefaktów w obrazie (np. nieregularnych krawędzi), z analizą dźwięku (nietypowe zniekształcenia głosu) oraz weryfikacją autentyczności emocji ujawnianych przez mimikę[62]. Dzięki temu można namierzyć niespójności między ruchem ust a wypowiadanym komunikatem

---

58 Zob. Y. Zhou, S.-N. Lim, Joint audio-visual deepfake detection, IEEE ICCV, 2021.

59 Zob. P. Yadav, A. Sharma, S. Jadon, S. Murala, DeepFake Detection using InceptionResNetV2 and LSTM, International Conference on Emerging Technologies, 2021.

60 Zob. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, Generative Adversarial Nets, Advances in Neural Information Processing Systems, 27, 2672-2680.

61 Zob. Y. Li, C.-C. J. Kuo, Video Content Analysis Using Multimodal Information, Springer, Berlin 2003.

62 Zob. D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), 2018.

czy też niepasujące do danego kontekstu modulacje głosu[63]. Co więcej, zestawienie wszystkich dostępnych modalności z danymi na temat źródła nagrania (np. strona internetowa czy profil o krótkim czasie funkcjonowania) dodatkowo wzmacnia pewność osądu o potencjalnej fałszywości przekazu[64]. Multimodalna analiza podlega jednak istotnym wyzwaniom metodologicznym. Po pierwsze, napotyka problem skalowalności i złożoności danych. Lawinowo rosnąca liczba nagrań wideo, infografik i memów w mediach społecznościowych czyni manualne weryfikowanie przekazów praktycznie niemożliwym. W odpowiedzi na tę sytuację opracowywane są półautomatyczne i automatyczne rozwiązania oparte na uczeniu maszynowym, które pozwalają na szybki przesiew dużych wolumenów treści. Po drugie, analiza utrudniona jest przez różnorodność językową i kulturową, ponieważ wiele narzędzi trenuje się na danych anglojęzycznych, co okazuje się nieadekwatne w badaniach przekazów funkcjonujących w odmiennych kręgach Z kolei subtelności takie jak ironia czy aluzje kulturowe wciąż pozostają trudne do uchwycenia przez systemy sztucznej inteligencji[65].

Kolejny aspekt dotyczy integracji różnorodnych modalności i wypracowania mechanizmów fuzji danych – tak, aby wykryć niespójności między obrazem, dźwiękiem a warstwą tekstową[66]. W obszarze technologii analizy *deepfake* stosuje się na przykład konwolucyjne sieci neuronowe (ang. *Convolutional Neural Networks*, CNN) do rozpoznawania obiektów w obrazie czy ruchów warg, a jednocześnie sieci rekurencyjne (ang. *Recurrent Neural Networks,* RNN), (ang. *Long Short-Term Memory Neural Network*, LSTM) do analizy warstwy dźwiękowej, w tym intonacji, rytmiki mowy czy potencjalnych cyfrowych zniekształceń[67]. Dobór właściwych cech do analizowania jest tu kluczowy, a co za tym idzie – konieczna jest interdyscyplinarna

---

[63] Zob. D. A. Sultan, L. M. Ibrahim, Video-audio multimodal for deepfake detection, Proceedings of the National Academy of Sciences, 2023.

[64] Zob. F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes, IEEE WACVW, 2019, s. 83-92.

[65] Zob. M. Kubit, Rozwój sztucznej inteligencji w świetle prawa Unii Europejskiej – dylematy i wyzwania, Think & Make, Warszawa, 2021.

[66] Zob. P. Yadav, A. Sharma, S. Jadon, S. Murala, DeepFake Detection using InceptionResNetV2 and LSTM, International Conference on Emerging Technologies, 2021.

[67] Zob. D. Güera and E. J. Delp, Deepfake video detection using recurrent neural networks,2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), 2018.

kooperacja specjalistów w dziedzinie informatyki, psychologii, bezpieczeństwa oraz komunikacji społecznej.

Rozwiązania z zakresu fuzji modalnej umożliwiają m.in. wspólne przetwarzanie wyników detekcji artefaktów wizualnych i nienaturalnych dźwięków[68]. Dzięki temu zwiększa się precyzja wychwytywania manipulacji, zwłaszcza gdy zarówno obraz, jak i audio noszą ślady sztucznej ingerencji. Dodatkowo, automatyczne rozpoznawanie emocji może wykazać nienaturalną ekspresję mimiki lub przesadne modulacje głosu, co również stanowi sygnał ostrzegawczy. W praktyce takie podejście pozwala analitykom *fact-checkingowym* wychwycić nawet sprytnie zamaskowane fałszerstwa. Narzędzia te znajdują zastosowanie w pracy profesjonalnych weryfikatorów informacji i służb odpowiedzialnych za bezpieczeństwo informacyjne. Umożliwiają one tzw. szybki przesiew materiałów potencjalnie fałszywych, kierując analityków do najbardziej podejrzanych treści. Niemniej implementacja rozwiązań multimodalnych napotyka liczne trudności organizacyjne: konieczne jest magazynowanie ogromnych zbiorów danych, zapewnienie wystarczającej mocy obliczeniowej i wdrażanie złożonych modeli sztucznej inteligencji. Poważnym problemem pozostaje też aspekt etyczny. Z jednej strony systemy te przetwarzają dane biometryczne (np. wizerunek, głos), co budzi obawy o nadmierną inwigilację, a z drugiej – brak odpowiednich narzędzi analizy wystawia odbiorców na masowe działania dezinformacyjne. Sytuacji nie ułatwia fakt, że manipulacje dotykają różnych kodów kulturowych. Symbole, memy czy narracje, które w jednym regionie uchodzą za neutralne, w innym mogą mieć silne konotacje polityczne bądź religijne. Istotnym problemem jest również wyjaśnialność algorytmów, czyli umiejętność zrozumienia, dlaczego system ocenił dany materiał jako fałszywy lub prawdziwy. Modele uczenia głębokiego działają często jako tzw. czarna skrzynka, co utrudnia ocenę ich trafności i weryfikację ewentualnych błędów. Równocześnie pomyłki w detekcji dezinformacji mogą być społecznie kosztowne – np. niesłuszne oznaczenie prawdziwego wideo jako fałszu bądź zaniechanie reakcji w sytuacji, gdy rozpowszechniane treści wywołują panikę bądź eskalują konflikt polityczny.

Mimo wszystkich wymienionych trudności analitycy i badacze wskazują, że w obliczu narastającej popularności sztucznie generowanych treści,

[68] Zob. R. Reimao, V. Tzerpos, Synthetic speech detection using neural networks, IEEE SpeD, 2021.

szczególnie *deepfake*, właśnie podejście multimodalne zapewnia najwyższą efektywność wykrywania zagrożeń. Tylko łączna weryfikacja obrazu, dźwięku i tekstu (oraz dostępnych informacji kontekstowych) umożliwia bowiem zdemaskowanie złożonych manipulacji, w których prostsze metody skoncentrowane na pojedynczej modalności mogłyby zawieść. Wykorzystanie takich rozwiązań wymaga współpracy środowisk naukowych, sektora IT i instytucji odpowiedzialnych za bezpieczeństwo, aby zadbać nie tylko o rozwój technologii, ale także o stworzenie odpowiednich uwarunkowań prawnych i etycznych. Wraz z dalszym postępem metod generatywnych rośnie bowiem znaczenie transparentności i zaufania społecznego do stosowanych systemów.

Dotychczasowe doświadczenia wskazują, że analiza multimodalna w przyszłości będzie się jeszcze intensywniej rozwijać. Jej skuteczność wzrośnie wraz z udoskonalaniem algorytmów *deep learning*, tworzeniem specjalistycznych baz danych i standardów wymiany informacji o zidentyfikowanych fałszerstwach. Wymaga to jednak nie tylko modernizacji rozwiązań technicznych, ale i wypracowania międzynarodowych zasad etycznego stosowania takich technologii. Tylko w ten sposób można zagwarantować, że badania nad multimodalnością i ich praktyczna implementacja nie zagrożą podstawowym prawom i wolnościom obywatelskim, chroniąc jednocześnie przestrzeń informacyjną przed coraz bardziej złożonymi formami manipulacji.

3.3. Wpływ technologii cyfrowych na bezpieczeństwo informacyjne państwa

Współczesne państwa funkcjonują w coraz bardziej złożonym ekosystemie informacyjnym, w którym technologie cyfrowe odgrywają rolę kluczowego katalizatora zmian społecznych, politycznych i gospodarczych. Postępująca digitalizacja sprawia, że zagadnienia związane z bezpieczeństwem informacyjnym nabierają bezprecedensowego znaczenia – szczególnie w kontekście zagrożeń płynących z manipulacji treściami audiowizualnymi oraz masowego przetwarzania danych. Celem niniejszego rozdziału jest ukazanie wielowymiarowego wpływu technologii cyfrowych na bezpieczeństwo informacyjne państwa, wskazanie najważniejszych wyzwań oraz omówienie istniejących i potencjalnych rozwiązań przeciwdziałających zagrożeniom.

Dynamiczny rozwój sztucznej inteligencji (ang. *artificial intelligence,* AI) oraz metod głębokiego uczenia (ang. *deep learning*) istotnie wpłynął na sposoby monitorowania i ochrony przestrzeni informacyjnej[69]. Jednocześnie rozwój nowych technologii przynosi ze sobą także wyzwania etyczne, prawne i społeczne, związane chociażby z ochroną danych osobowych, transparentnością algorytmów oraz ryzykiem nadużyć przy wykorzystaniu technologii głębokiego uczenia. W perspektywie strategicznej, rozwój sztucznej inteligencji wymaga tworzenia odpowiednich standardów i mechanizmów kontroli, które umożliwią korzystanie z jej dobrodziejstw, ograniczając potencjalne zagrożenia, mogące mieć destrukcyjny wpływ na bezpieczeństwo państwa i społeczeństwa.

Zaawansowane algorytmy umożliwiają wykrywanie i analizę różnorodnych materiałów audiowizualnych w niemal rzeczywistym czasie, co znacząco podnosi efektywność walki z zagrożeniami informacyjnymi[70]. Dynamiczny rozwój sztucznej inteligencji oraz metod głębokiego uczenia istotnie wpłynął na sposoby monitorowania i ochrony przestrzeni informacyjnej, stając się integralnym elementem globalnej rewolucji informacyjnej, określanej również mianem czwartej rewolucji przemysłowej. Rozwiązania te obejmują między innymi rozpoznawanie obrazów i twarzy, realizowane przy użyciu systemów bazujących na sieciach neuronowych. Pozwalają one z dużą dokładnością identyfikować osoby występujące w materiałach audiowizualnych, co jest istotnym narzędziem w zwalczaniu terroryzmu czy przestępczości zorganizowanej[71]. Wykorzystuje się również oprogramowanie do analizy mowy i emocji oparte na przetwarzaniu języka naturalnego (ang. *Natural Language Processing*, NLP). Systemy te są zdolne nie tylko do automatycznej transkrypcji nagrań, lecz także do oceny ich kontekstu oraz emocjonalnego nacechowania. W połączeniu z analizą sentymentu wypowiedzi w mediach społecznościowych technologia ta znacznie ułatwia identyfikowanie sytuacji prowadzących do eskalacji agresji lub szerzenia mowy

[69] Zob. A. Lasota-Kapczuk, Sztuczna inteligencja w kształtowaniu bezpieczeństwa państwa w II dekadzie XXI wieku, Rozprawa doktorska, Uniwersytet Przyrodniczo-Humanistyczny w Siedlcach, 2021.

[70] Zob. T. Wróblewski, Sztuczna inteligencja jako narzędzie do walki z dezinformacją, Poznań, IJOLS 2024, 17 (1), 2024, s. 149-166.

[71] Zob. Zob. N. Dakhil, A.M. Abdulazeez, Face recognition based on deep learning: A comprehensive review, Indonesian Journal of Computer Science, 13(3), 2024.

nienawiści[72]. Szczególną rolę odgrywają także systemy wczesnego ostrzegania, polegające na monitorowaniu nagłych wzrostów liczby wyświetleń czy udostępnień materiałów podejrzanych o rozpowszechnianie dezinformacji, co pozwala służbom państwowym szybko identyfikować zagrożenia i odpowiednio wcześnie blokować rozprzestrzenianie się szkodliwych treści[73]. Wszystkie wymienione narzędzia znacząco poszerzają możliwości operacyjne państw w zakresie prewencji i neutralizacji zagrożeń o charakterze informacyjnym[74]. Coraz częściej są one także integrowane z systemami analizy big data, co umożliwia odkrywanie korelacji oraz wzorców zachowań z ogromnych zbiorów danych pochodzących z wielu źródeł – mediów społecznościowych, rejestrów publicznych czy baz danych operatorów telekomunikacyjnych[75].

Analiza *big data* opiera się na integracji różnorodnych źródeł danych, takich jak posty w mediach społecznościowych, artykuły prasowe, nagrania wideo, rejestry transakcji czy informacje geolokalizacyjne[76,77]. Celem takiego podejścia jest:

- Identyfikacja nietypowych wzorców interakcji – dzięki metodom statystycznym i uczeniu maszynowemu możliwe jest wykrycie anomalii w funkcjonowaniu sieci społecznościowych, np. nagłych „pików" aktywności wokół określonego tematu czy obecności powtarzalnych komunikatów o jednakowej treści.
- Mapowanie kluczowych węzłów – poprzez analizę sieciową (ang. *Social Network Analysis*, SNA) można określić najważniejsze konta i grupy odpowiedzialne za rozprzestrzenianie treści dezinformacyjnych. Pozwala to na szybkie namierzenie podmiotów

[72] Zob. J. Kowalski, Wykorzystanie analizy mowy w badaniach nad bezpieczeństwem informacyjnym, Studia nad Komunikacją Społeczną, 7, 112-125.

[73] Zob. L. Nguyen, M. Chen, Early Detection of Cyber Threats: Big Data Analytics Approaches, Cyber Security International Journal, 2, 79-95.

[74] Zob. H. Huang, J. Li, Research on Network Communication Model and Network Security Technology through Big Data, IEEE, 2021.

[75] Zob. H. Batorowska, P. Motylińska, O. Wasiuta, Bezpieczeństwo informacyjne i medialne w czasach nadprodukcji informacji, Wydawnictwo Naukowe i Edukacyjne Stowarzyszenia Bibliotekarzy Polskich, 2020.

[76] Zob. M. Tabakow, J. Korczak, B. Franczyk, Big Data – definicje, wyzwania i technologie informatyczne. Informatyka Ekonomiczna, 1(31).

[77] Zob. M. Dehmer, S. Borgert, F. Emmert-Streib, Big Data Analysis in Modern Information Warfare, European Journal of Security Studies, 4, 14-27.

kierujących kampaniami manipulacyjnymi i ich zablokowanie bądź zneutralizowanie[78].

- Analizę semantyczną i kontekstową – zaawansowane metody przetwarzania języka naturalnego ułatwiają wyodrębnienie słów kluczowych, wzorców narracji czy emocjonalnego zabarwienia przekazów, co pomaga ustalić, czy mamy do czynienia z działaniami koordynowanymi i w jakim stopniu realnie wpływają one na opinię publiczną.

Państwa coraz częściej korzystają z usług firm specjalizujących się w analizie big data, które dostarczają platformy i narzędzia do automatycznego monitorowania treści w Internecie. Z jednej strony pozwala to na efektywne zarządzanie ryzykiem informacyjnym, z drugiej jednak rodzi poważne wątpliwości dotyczące ochrony prywatności i wolności słowa. Dlatego też priorytetem staje się wprowadzanie transparentnych procedur i proporcjonalnych rozwiązań prawnych, które z jednej strony umożliwiają skuteczną reakcję na zagrożenia, a z drugiej chronią prawa obywateli.

Jednym z najbardziej niepokojących przykładów zagrożeń wynikających z rozwoju współczesnych technologii cyfrowych są *deepfake'i* – sztucznie generowane materiały audiowizualne, które w realistyczny sposób naśladują głos oraz wizerunek określonych osób. Dzięki wykorzystaniu zaawansowanych technik uczenia maszynowego, szczególnie generatywnych sieci neuronowych (ang. *Generative Adversarial Networks*, GAN), możliwe jest obecnie tworzenie materiałów, które są trudne do odróżnienia od prawdziwych nagrań audio lub wideo. Zjawisko to niesie ze sobą istotne zagrożenia dla bezpieczeństwa informacyjnego oraz prywatności, umożliwiając manipulowanie opinią publiczną poprzez fałszywe wypowiedzi przypisywane politykom czy innym ważnym osobistościom[79]. *Deepfake'i* mogą być wykorzystywane do wywoływania chaosu informacyjnego, szantażu lub dyskredytacji osób publicznych, co negatywnie wpływa na stabilność społeczną i polityczną. Skuteczność takich manipulacji wynika z szybkiego rozprzestrzeniania się treści w mediach społecznościowych, co może prowadzić do eskalacji napięć społecznych lub międzynarodowych, zanim jeszcze

[78] Zob. P. Zakrzewski, Analiza sieci społecznych w badaniach nad bezpieczeństwem militarnym i niemilitarnym państwa, Zeszyty Naukowe WSOWL, 1(167), 47-56.

[79] Zob. H. Khalid, S. Tariq, M. S. U. Hassan, M. H. Rehmani, Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors, arXiv preprint, 2021.

manipulacja zostanie wykryta i zdemaskowana. Aby przeciwdziałać tego typu zagrożeniom, rozwijane są zaawansowane systemy automatycznego wykrywania fałszerstw. Opierają się one na analizie specyficznych artefaktów cyfrowych generowanych podczas tworzenia *deepfake'ów*, jak również na wykrywaniu nienaturalnych ruchów mimicznych twarzy czy anomalii zachowania występujących postaci[80]. Mimo tych wysiłków, wraz z dalszym rozwojem technologii, kolejne generacje *deepfake'ów* stają się coraz trudniejsze do rozpoznania, co powoduje konieczność ciągłego ulepszania narzędzi analitycznych i podejmowania coraz bardziej kompleksowych metod wykrywania. Szczególnie efektywne okazuje się podejście multimodalne, łączące jednoczesną analizę obrazu, dźwięku oraz tekstu, które umożliwia skuteczniejszą ochronę przestrzeni informacyjnej przed zagrożeniami generowanymi przez zaawansowane manipulacje audiowizualne.

Zaawansowanym uzupełnieniem metod opartych na analizie multimodalnej są technologie *blockchain*, czyli tzw. rozproszone rejestry (ang. *distributed ledgers*). Technologia ta umożliwia przechowywanie tzw. *hashy* (unikalnych podpisów cyfrowych) plików multimedialnych w publicznych bazach danych, co znacząco utrudnia późniejsze modyfikacje materiałów audiowizualnych bez ich wykrycia. Rozwiązanie to pozwala szybko i precyzyjnie potwierdzić autentyczność oryginalnego nagrania, a ewentualne późniejsze manipulacje są łatwe do wykrycia dzięki porównaniu aktualnego *hasha* z tym przechowywanym w łańcuchu bloków. Podobną filozofię reprezentują inicjatywy takie jak C2PA (ang. *Coalition for Content Provenance and Authenticity*), zrzeszające technologicznych gigantów takich jak Microsoft, Adobe czy BBC. Opracowują one specjalistyczne standardy techniczne umożliwiające automatyczne potwierdzanie źródła i integralności treści audiowizualnych, co może w przyszłości stać się kluczowym elementem systemów weryfikacji autentyczności materiałów w Internecie[81]. Jednocześnie, implementacja tego rodzaju technologii nie jest pozbawiona wyzwań. Wymaga ona szerokiej współpracy międzysektorowej – państw, firm technologicznych, mediów oraz organizacji społecznych – oraz ustanowienia odpowiedniej infrastruktury technicznej. Należy również pamiętać, że sama

[80] Zob. S. Agarwal, T. El-Gaaly, H. Farid, S.-N. Lim, Detecting Deep-Fake Videos from Appearance and Behavior, arXiv preprint, 2020.

[81] Coalition for Content Provenance and Authenticity (C2PA), C2PA Technical Specification, 2022.

integracja pliku multimedialnego z *blockchainem* nie uniemożliwia stworzenia odrębnego, zmanipulowanego materiału od podstaw. Rozwiązania tego rodzaju stanowią raczej narzędzie do potwierdzania integralności oryginalnych treści i monitorowania późniejszych zmian, niż metodę kompleksowo zapobiegającą powstawaniu nowych zmanipulowanych treści.

Ostatnim, lecz niezwykle ważnym aspektem wpływu technologii cyfrowych na bezpieczeństwo informacyjne państwa jest konieczność wypracowania adekwatnych regulacji prawnych oraz standardów etycznych, które będą odpowiadać na szybko zmieniające się zagrożenia[82]. Rosnące możliwości technologiczne w zakresie masowego monitorowania i analizy treści internetowych, choć skuteczne w identyfikacji i neutralizacji zagrożeń, niosą ze sobą poważne ryzyko naruszania prawa do prywatności oraz ograniczenia wolności słowa[83]. Dlatego tak istotne jest stosowanie zasady proporcjonalności, przejrzystości procedur, jak również poddanie działań instytucji państwowych skutecznej kontroli sądowej lub parlamentarnej. Coraz bardziej widoczna staje się także kwestia odpowiedzialności dużych platform internetowych za treści publikowane przez użytkowników. Przykładowo, wprowadzona w Unii Europejskiej regulacja *Digital Services Act* zobowiązuje właścicieli platform społecznościowych do monitorowania, usuwania i ograniczania zasięgu materiałów uznanych za szkodliwe lub niebezpieczne, co jednak budzi dyskusję o potencjalnym ograniczeniu wolności wypowiedzi i swobody dostępu do informacji[84]. Dodatkowo, poszczególne państwa rozwijają własne regulacje umożliwiające szybkie reagowanie na incydenty dezinformacyjne w sytuacjach kryzysowych, na przykład podczas kampanii wyborczych lub konfliktów. Mimo to nadal brakuje spójnej, międzynarodowej architektury regulacyjnej zdolnej odpowiedzieć na zagrożenia informacyjne o globalnym zasięgu.

Nie można także zapominać o kluczowej roli edukacji medialnej i informacyjnej społeczeństwa, bez której nawet najbardziej zaawansowane technologie czy skuteczne regulacje prawne mogą okazać się niewystarczające

---

[82] Zob. J. Wieczorkowski, Big data a prywatność. Naruszenie prywatności w świecie wirtualnym, Roczniki KAE, 45, 2017.

[83] Zob. J. Wieczorkowski, Metody big data w zapewnieniu bezpieczeństwa publicznego a problem prywatności, Roczniki KAE, nr 46, 2018.

[84] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689, w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji, Dziennik Urzędowy UE, 2024.

w walce z dezinformacją. Państwa powinny realizować szeroko zakrojone programy edukacyjne, których celem jest rozwijanie kompetencji krytycznego odbioru treści, umiejętności rozpoznawania manipulacji oraz świadomego i odpowiedzialnego korzystania ze źródeł informacji. W rezultacie tylko zintegrowane podejście, obejmujące rozwój technologiczny, odpowiednie regulacje prawne, standardy etyczne oraz systematyczną edukację obywateli, może zagwarantować efektywne zabezpieczenie przestrzeni informacyjnej państwa przed coraz bardziej wyrafinowanymi formami zagrożeń cyfrowych.

Podsumowując, technologie cyfrowe są obecnie podstawą rozwoju społeczno-gospodarczego, lecz jednocześnie generują szereg wyzwań i zagrożeń dla bezpieczeństwa informacyjnego państwa. Narzędzia oparte na sztucznej inteligencji, analizie big data oraz technologii *blockchain* znacząco zwiększają zdolność państw do identyfikacji, przeciwdziałania i neutralizacji manipulacji informacyjnych, takich jak *deepfake'i* czy kampanie dezinformacyjne. Z drugiej jednak strony, postęp technologiczny umożliwia stosowanie coraz bardziej zaawansowanych form manipulacji, którym coraz trudniej przeciwdziałać. W związku z tym skuteczna strategia bezpieczeństwa informacyjnego wymaga jednoczesnego wdrażania zaawansowanych technologii wykrywających manipulacje, rozwoju adekwatnych regulacji prawnych oraz zwiększania społecznej świadomości zagrożeń. Kluczowe staje się także znalezienie odpowiedniej równowagi między ochroną przestrzeni informacyjnej a poszanowaniem podstawowych praw obywatelskich. Tylko dzięki zintegrowanym, międzynarodowym działaniom obejmującym współpracę sektora publicznego, prywatnego i społeczeństwa możliwe będzie skuteczne przeciwdziałanie zagrożeniom cyfrowym i zapewnienie integralności przestrzeni informacyjnej w dobie dynamicznych przemian technologicznych.

3.4. Koncepcja wojny informacyjnej w kontekście audiowizualnym

W literaturze przedmiotu wskazuje się, że wojna informacyjna (ang. *information warfare*, IW) jest specyficznym rodzajem rywalizacji, w której obiektem ataku staje się sfera informacyjno-psychologiczna społeczeństwa,

a głównym celem jest narzucenie mu określonego obrazu rzeczywistości[85]. Inaczej mówiąc, agresor dąży do ukształtowania percepcji i emocji odbiorców w pożądany dla siebie sposób, osłabiając jednocześnie wolę oporu i zdolność krytycznego myślenia w zaatakowanym społeczeństwie. Strategiczne cele wojny informacyjnej koncentrują się na destabilizacji struktur informacyjnych przeciwnika oraz dezorientacji jego społeczeństwa i aparatu decyzyjnego. Głównym zadaniem takich działań jest celowe fałszowanie, zniekształcanie informacji oraz obniżanie morale przeciwnika w celu uzyskania nad nim przewagi[86]. Ostatecznie chodzi o osiągnięcie własnych celów politycznych lub militarnych bez otwartej konfrontacji zbrojnej – poprzez wywarcie wpływu na postawy, przekonania i decyzje podejmowane przez wrogie państwo lub grupę. Obiektem ataku może być zarówno społeczeństwo (ludność cywilna), jak i siły zbrojne danego państwa. W odróżnieniu od klasycznej propagandy, która może przybierać formy jawne i nie zawsze wiąże się z intencją wrogiego działania, wojna informacyjna oznacza zorganizowane, długofalowe kampanie dezinformacyjne realizujące obce cele strategiczne kosztem zaatakowanego podmiotu. Działania te prowadzone są wielowymiarowo – w sferze politycznej, społecznej, ekonomicznej i oczywiście medialnej – w celu osłabienia przeciwnika oraz wzmocnienia własnej pozycji. W praktyce arsenał środków wykorzystywanych w wojnie informacyjnej jest bardzo szeroki. Obejmuje on zarówno techniki propagandowe (planowe, masowe oddziaływanie na psychikę poprzez bodźce informacyjne), jak i dezinformacyjne (świadome rozpowszechnianie fałszywych lub wprowadzających w błąd komunikatów). Stosowane metody to m.in. rozpowszechnianie nieprawdziwych lub zmanipulowanych wiadomości, zalewanie przestrzeni informacyjnej sprzecznymi narracjami (tzw. biały szum informacyjny) czy prezentowanie faktów w stronniczym kontekście. Typowe jest też fabrykowanie dowodów w postaci materiałów medialnych, posługiwanie się fałszywymi tożsamościami w Internecie (trolle, boty) czy tworzenie całych sieci pseudo-mediów szerzących określoną narrację. Istotnym elementem bywa propaganda komputerowa polegająca na automatycznym generowaniu i dystrybucji przekazów z wykorzystaniem botów w mediach

[85] Zob. R. Waltzman, The Weaponization of Information: The Need for Cognitive Security, RAND Corporation, 2017.

[86] Zob. T. Darmohrai, Nowoczesna wojna informacyjna: metody i środki przeciwdziałania, Media i Społeczeństwo, 17/2022.

społecznościowych. Wszystkie te działania mają służyć dezintegracji informacyjnej przeciwnika – zniekształceniu jego obrazu świata, wywołaniu chaosu informacyjnego i podważeniu zaufania do oficjalnych komunikatów. W efekcie społeczeństwo zaatakowanego państwa może ulec dezinformacji na tyle, że będzie podejmować decyzje zgodne z interesem agresora, nie zdając sobie z tego sprawy. Co istotne, prowadzenie wojen informacyjnych jest relatywnie niskokosztowe i trudniejsze do jednoznacznego przypisania sprawcy niż konwencjonalne działania militarne. To sprawia, że stanowią one atrakcyjne narzędzie realizacji celów geopolitycznych w XXI wieku.

Treści audiowizualne w operacjach informacyjnych, dezinformacyjnych i propagandowych – obejmujące obraz (foto, wideo) oraz dźwięk – odgrywają coraz istotniejszą rolę w prowadzeniu wojen informacyjnych. Wynika to z faktu, że komunikaty wizualne i dźwiękowe silniej oddziałują na emocje i uwagę odbiorców niż komunikaty czysto tekstowe. Obraz propagandowy wykorzystywany jest w konfliktach od dekad (przykładem mogą być plakaty i kroniki filmowe z okresu II wojny światowej czy audycje radiowo-telewizyjne w czasie zimnej wojny), jednak rozwój telewizji satelitarnej, a następnie Internetu i mediów społecznościowych, zwielokrotnił zasięg i szybkość oddziaływania takich treści. Wizualne materiały propagandowe stały się centralnym komponentem działań ekstremistycznych i terrorystycznych, przy czym nowoczesne technologie pozwalają tym grupom produkować filmy o jakości niemal dorównującej profesjonalnym mediom oraz natychmiast dystrybuować je globalnie za pośrednictwem sieci[87]. Jednocześnie podkreślana jest powszechna w naukach społecznych zgodność co do potężniejszej siły oddziaływania obrazu w porównaniu ze słowem pisanym, co ma istotne konsekwencje strategiczne. Innymi słowy, umiejętnie przygotowane treści audiowizualne mogą skuteczniej kształtować opinie i postawy niż tradycyjne artykuły czy odezwy, zwłaszcza w społeczeństwie informacyjnym skłonnym do percepcji poprzez media elektroniczne. W ramach operacji informacyjnych propaganda audiowizualna przybiera rozmaite formy. Mogą to być oficjalne przekazy rządowe: filmy dokumentalne, wystąpienia przywódców państw adresowane do zagranicznej opinii publicznej oraz materiały tworzone przez organizacje radykalne jak nagrania propagandowe organizacji terrorystycznych służące rekrutacji

[87] Zob. C.E. Dauber, C.K. Winkler, Visual Propaganda and Extremism in the Online Environment, Carlisle Barracks, PA: U.S. Army War College Press, 2014.

i zastraszaniu przeciwników, czy pozornie oddolne treści krążące w mediach społecznościowych jak *viralowe* filmiki i *memy* o zabarwieniu politycznym. Klasyczne przykłady stanowią tu nagrania wideo publikowane przez tzw. Państwo Islamskie, które były profesjonalnie montowane i stylizowane na przekaz informacyjny, ale miały na celu szerzenie ekstremistycznej ideologii i demonstrację okrucieństwa w celu zastraszenia odbiorców. Tego typu treści, często rozpowszechniane poprzez media społecznościowe i platformy typu YouTube czy TikTok, potrafią dotrzeć do milionów osób w krótkim czasie, kształtując ich postrzeganie konfliktów i zjawisk społeczno-politycznych. Szczególną odmianą audiowizualnych operacji dezinformacyjnych są materiały wykorzystujące manipulacje cyfrowe obrazu i dźwięku. Zaliczają się do nich zarówno relatywnie proste przeróbki (tzw. *cheapfakes*, np. wyrwanie wypowiedzi z kontekstu, montaż wprowadzający w błąd, fotomontaże), jak i wysoce zaawansowane techniki z użyciem sztucznej inteligencji, takie jak *deepfake*. Technologia *deepfake* umożliwia generowanie fałszywych nagrań wideo, w których np. twarz i głos danej osoby (polityka, celebryty) zostają syntetycznie podmienione, sprawiając wrażenie autentycznego wystąpienia. W rękach sprawnego propagandysty taki materiał może posłużyć do wywołania sensacji lub kryzysu – przykładowo poprzez wygenerowanie wypowiedzi, której dana osoba nigdy nie wygłosiła. Audiowizualne fałszerstwa mogą bowiem nie tylko wprowadzać odbiorców w błąd, ale wręcz podważać zaufanie społeczeństwa do materiałów filmowych jako dowodów. Jeśli obywatele zaczną wątpić w autentyczność wszelkich przekazów wideo (zasadnie lub nie), łatwiej szerzyć chaos informacyjny i podkopać fundamenty debaty publicznej opartej na faktach.

Rozwój technologii cyfrowych wywarł dwojaki wpływ na zjawisko wojny informacyjnej. Z jednej strony, nowoczesne media elektroniczne stały się głównym polem bitwy informacyjnej – to w cyberprzestrzeni dochodzi dziś do większości operacji propagandowych i dezinformacyjnych. Internet umożliwia agresorom błyskawiczne i globalne rozpowszechnianie treści audiowizualnych o manipulacyjnym charakterze przy minimalnych kosztach infrastrukturalnych[88]. Media społecznościowe sprzyjają *viralowemu* rozchodzeniu się *fake newsów*, a algorytmy personalizujące treści mogą mimowolnie tworzyć odbiorcom tzw. bańki informacyjne, w których dezinformacja

[88] Zob. Z. Ciekanowski, Cyberprzestrzeń jako obszar współczesnych zagrożeń, Społeczeństwo jako uczestnik bezpieczeństwa, Akademia Sztuki Wojennej, 2021, s. 7-20.

pozostaje niekwestionowana. Ponadto narzędzia typu *botnet* pozwalają automatycznie generować i dystrybuować setki tysięcy wpisów, komentarzy czy multimediów wzmacniających wybraną narrację, co sztucznie zwiększa jej zasięg. W efekcie skala i szybkość zalewu informacyjnego przerasta możliwości tradycyjnych instytucji państwowych w zakresie monitorowania i reagowania. W mediach społecznościowych publikowane są niezliczone krótkie klipy wideo i transmisje na żywo. Taka lawinowa podaż treści powoduje, że wychwycenie w realnym czasie potencjalnie groźnych przekazów podżegających do przemocy lub zawierających dezinformacje mogące wywołać panikę, jest ogromnym wyzwaniem dla służb bezpieczeństwa. Z drugiej strony, rozwój technologiczny dostarcza jednak także środków obrony przed zagrożeniami informacyjnymi. Współczesne systemy bezpieczeństwa państwa coraz częściej wykorzystują zaawansowane narzędzia analityczne oparte na sztucznej inteligencji i uczeniu maszynowym, zdolne do automatycznego przetwarzania i analizy wielkich zbiorów danych multimedialnych. Kluczowym kierunkiem jest tutaj analiza multimodalna, łącząca informacje z różnych kanałów jednocześnie – obrazu, dźwięku, a nierzadko także warstwy tekstowej (np. rozpoznanej mowy czy napisów towarzyszących materiałowi). Tego typu podejście pozwala efektywnie wykrywać nawet złożone formy manipulacji. Jak pokazują najnowsze badania, systemy detekcji oparte na fuzji danych audiowizualnych potrafią skuteczniej identyfikować fałszywe nagrania niż analizy skupione wyłącznie na jednej modalności[89]. Przykładowo, algorytm analizujący jednocześnie mimikę postaci na filmie oraz modulację jej głosu ma większą szansę wykryć niespójności typowe dla *deepfake* (np. drobne artefakty wizualne lub nienaturalne brzmienie poszczególnych słów), które mogłyby umknąć przy badaniu tylko obrazu lub tylko dźwięku. Tylko łączna weryfikacja wielu aspektów przekazu umożliwia zdemaskowanie najbardziej zaawansowanych falsyfikatów. Co więcej, narzędzia analityczne coraz częściej uwzględniają kontekstowe metadane takie jak źródło pochodzenia pliku, daty modyfikacji, czy geolokalizacja oraz porównawcze baz danych autentycznych materiałów, aby ocenić wiarygodność nowych treści pojawiających się w sieci. Ważnym elementem przeciwdziałania dezinformacji jest również rozwój cyfrowej higieny informacyjnej i standaryzacja procedur weryfikacji informacji. Pojawiają

[89] Zob. D. Salvi, H. Liu, S. Mandelli, P. Bestagini, W. Zhou, W. Zhang, S. Tubaro, A Robust Approach to Multimodal Deepfake Detection, J Imaging, 2023.

się inicjatywy tworzenia systemów ostrzegawczych zdolnych do wychwytywania anomalii w ekosystemie informacyjnym, nienaturalnego nagłego wzrostu popularności określonego filmu czy *hashtagu*, co może wskazywać na skoordynowaną akcję propagandową. Organizacje międzynarodowe i instytucje państwowe opracowują także wytyczne odnośnie uwierzytelniania materiałów audiowizualnych. Jest to praktykowane poprzez znakowanie plików wideo unikalnymi znacznikami kryptograficznymi w momencie ich tworzenia, co umożliwi późniejsze sprawdzenie, czy dany plik został zmodyfikowany. Równolegle prowadzone są szeroko zakrojone szkolenia z zakresu bezpieczeństwa informacyjnego oraz kampanie na rzecz edukacji medialnej społeczeństwa, aby podnieść ogólną odporność odbiorców na dezinformację. Jak stwierdzono w raporcie niezależnej grupy ekspertów dla Komisji Europejskiej, dezinformacja – definiowana jako celowo fałszywe lub wprowadzające w błąd informacje rozpowszechniane dla zysku lub świadomego wprowadzenia odbiorców w błąd – stała się bronią w arsenale współczesnych wojen informacyjnych[90]. W odpowiedzi konieczne jest łączenie najnowszych zdobyczy techniki z działaniami organizacyjnymi i edukacyjnymi. Tylko całościowe, wielowymiarowe podejście pozwoli skutecznie chronić przestrzeń informacyjną państwa przed coraz bardziej wyrafinowanymi formami manipulacji audiowizualnej, zachowując równocześnie podstawowe wartości demokratyczne i społeczne oparte na dostępie do rzetelnej informacji.

[90] Komisja Europejska, A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation, 2018.

IV

Współczesne zagrożenia dezinformacyjne i medialne

4.1. Rola platform cyfrowych i algorytmów w rozpowszechnianiu dezinformacji

Współczesna przestrzeń informacyjna jest w coraz większym stopniu modelowana przez platformy cyfrowe, których mechanizmy działania opierają się na złożonych algorytmach rekomendujących treści użytkownikom. Meta (Facebook, Instagram), X (dawny Twitter), YouTube, TikTok i inne serwisy społecznościowe nie tylko pośredniczą w komunikacji i wymianie informacji, ale aktywnie uczestniczą w selekcji, promowaniu i kształtowaniu treści docierających do użytkowników. W tym kontekście coraz więcej uwagi poświęca się analizie ich roli w rozpowszechnianiu dezinformacji, manipulacji medialnych i radykalizacji postaw społecznych.

Algorytmy stosowane przez platformy społecznościowe mają na celu przede wszystkim maksymalizację zaangażowania użytkownika. Treści są dobierane na podstawie wcześniejszych interakcji, preferencji, czasu spędzonego nad daną treścią czy rodzaju reakcji (polubienia, komentarze, udostępnienia). Mechanizmy te, choć skuteczne z punktu widzenia utrzymania uwagi użytkownika, prowadzą często do tzw. efektu „komory echa" (echo chamber), w której jednostka styka się wyłącznie z informacjami potwierdzającymi jej wcześniejsze przekonania[91]. W ten sposób zanika różnorodność poglądów, a użytkownicy są stopniowo izolowani od odmiennych

[91] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, The echo chamber effect on social media, Proc. Natl. Acad. Sci. U.S.A., 118(9), 2021.

narracji. Zjawisko to sprzyja polaryzacji społecznej, radykalizacji przekonań oraz wzrostowi podatności na dezinformację.

Kolejnym problemem związanym z działaniem algorytmów jest ich preferencja dla treści wywołujących silne reakcje emocjonalne – sensacyjnych, kontrowersyjnych, prowokacyjnych. Badania pokazują, że fałszywe informacje często rozprzestrzeniają się szybciej niż prawdziwe, właśnie ze względu na ich atrakcyjność emocjonalną i zdolność do wywoływania natychmiastowych reakcji. W konsekwencji, algorytmy promujące zaangażowanie mogą nieświadomie preferować treści dezinformacyjne, wzmacniając ich zasięg i wpływ społeczny[92]. W kontekście bezpieczeństwa informacyjnego staje się to szczególnie groźne w okresach kryzysów, kampanii wyborczych, konfliktów zbrojnych czy epidemii – czyli wtedy, gdy społeczeństwo jest najbardziej podatne na manipulację i dezinformację.

Odpowiedzialność platform cyfrowych za treści rozpowszechniane przez użytkowników stanowi jeden z najbardziej kontrowersyjnych tematów współczesnej debaty publicznej i regulacyjnej. Przez wiele lat operatorzy platform internetowych korzystali z uprzywilejowanego statusu pośrednika – nie ponosząc odpowiedzialności za treści publikowane przez użytkowników. Jednak coraz częściej państwa oraz organizacje międzynarodowe (np. UE) podnoszą konieczność uregulowania tej kwestii. Pojawiają się inicjatywy legislacyjne, takie jak europejski Akt o usługach cyfrowych (DSA), które mają zobowiązać platformy do szybszego reagowania na szkodliwe treści, większej przejrzystości algorytmów oraz skuteczniejszej moderacji[93].

W praktyce jednak działania platform cyfrowych wobec dezinformacji są nierówne, często spóźnione, a ich skuteczność budzi wątpliwości. Przykładowo, podczas pandemii COVID-19 niektóre serwisy (np. YouTube, Facebook) podejmowały próby flagowania lub usuwania treści dezinformujących na temat szczepionek czy przebiegu choroby, jednak równocześnie inne kanały z powodzeniem szerzyły fake newsy wśród milionów odbiorców[94].

---

[92] K. Jędrasiak, Analiza cyberzagrożeń współczesnych kanałów komunikacyjnych [w:] Horyzonty sztucznej inteligencji a Przemysł 5.0, Z. Dacko-Pikiewicz, K. Szczepańska-Woszczyna, M. Lis (red.), Akademia WSB, Dąbrowa Górnicza 2025, s. 39.

[93] J. Doniec-Niezgoda, Zaprzepaszczone szanse Aktu o Usługach Cyfrowych? – rozważania dotyczące definicji platformy internetowej – część I, Prawo Mediów Elektronicznych. Kwartalnik Naukowy, 1, 2024, 20-21.

[94] S. Wigienka, Przeciwdziałanie dezinformacji w dobie pandemii COVID-19, Com.press, 1, 2021, s. 129.

Podobna sytuacja miała miejsce w kontekście wojny w Ukrainie, kiedy to rosyjska propaganda wideo z wykorzystaniem materiałów audiowizualnych była masowo udostępniana i komentowana mimo wysiłków platform do jej ograniczania[95]. Platformy niejednokrotnie deklarowały walkę z dezinformacją, lecz ich działania pozostawały ograniczone do form symbolicznych lub punktowych interwencji, bez systemowego podejścia.

Nie bez znaczenia są również ograniczenia moderacji treści, zwłaszcza w skali globalnej. Automatyzacja tego procesu, oparta na sztucznej inteligencji i uczeniu maszynowym, często nie radzi sobie z kontekstem kulturowym, ironią, językiem wizualnym czy subtelną manipulacją. Z kolei moderacja ręczna, prowadzona przez zespoły ludzi, jest kosztowna, czasochłonna i obciążona ryzykiem błędów. Dodatkowo, moderacja budzi napięcia związane z wolnością słowa, cenzurą i arbitralnością decyzji podejmowanych przez prywatne korporacje, które nie są podmiotami demokratycznej kontroli.

Zjawisko wpływu platform cyfrowych na rozprzestrzenianie dezinformacji stawia przed państwami i społeczeństwami poważne wyzwania. Wymaga ono z jednej strony rozwoju odpowiednich ram prawnych, które zapewnią przejrzystość algorytmów i zwiększą odpowiedzialność operatorów platform, z drugiej zaś – inwestycji w edukację medialną, świadomość użytkowników i niezależne dziennikarstwo. Tylko zrównoważone podejście – łączące innowację, bezpieczeństwo i wolność wypowiedzi – pozwoli zbudować odporną przestrzeń informacyjną, w której technologia nie będzie sprzyjać dezinformacji, lecz stanie się narzędziem jej ograniczania.

4.2. *Fake newsy* i mechanizmy manipulacji audiowizualnej

Współczesne państwa coraz wyraźniej dostrzegają, że ich bezpieczeństwo narodowe nie ogranicza się jedynie do sfery militarnej czy ekonomicznej. W dobie społeczeństwa informacyjnego, gdzie wiadomości rozpowszechniane są z ogromną szybkością i w niemal nieograniczonym zasięgu, kluczową rolę zaczyna odgrywać bezpieczeństwo informacyjne. Jednym z najpoważniejszych zagrożeń w tym obszarze jest zjawisko dezinformacji – umyślne, celowe

[95] J. Czerwiński, Pożyteczni idioci" kampanii dezinformacyjnych. Mechanizmy tworzenia się społeczności wirtualnych rozpowszechniających fałsz i manipulację w sieci, Konteksty Społeczne, 2(20), 2022, s. 6

rozpowszechnianie fałszywych wiadomości lub przekazów wprowadzających w błąd. Szczególnie niebezpieczne okazują się treści w formie audiowizualnej, które potrafią silniej działać na emocje i szybciej wywoływać określone reakcje społeczne. W rezultacie zjawisko manipulacji audiowizualnych zyskało w ostatnich latach rangę istotnego wyzwania dla stabilności państw.

Termin *fake news* odnosi się do świadomie sfabrykowanych lub zmanipulowanych informacji przedstawianych w formacie wiadomości medialnych, których celem jest celowe wprowadzenie odbiorców w błąd[96]. *Fake newsy* obejmują zarówno całkowicie fikcyjne treści, jak również przekazy częściowo prawdziwe, które zostały zniekształcone przez wybiórcze prezentowanie faktów lub wyrwanie ich z pierwotnego kontekstu[97]. Pojęcie to zyskało popularność po wyborach prezydenckich w USA w 2016 roku, niemniej jednak praktyka świadomego fabrykowania informacji ma znacznie dłuższą historię, sięgającą początków prasy drukowanej. Już w XIX wieku pojawiały się przykłady tzw. „żółtego dziennikarstwa", które wykorzystywało sensacyjne treści w celu manipulacji opinią publiczną. Współcześnie problem *fake newsów* ewoluował wraz z postępem technologicznym, a ich zasięg i tempo rozpowszechniania znacząco wzrosły dzięki mediom cyfrowym takim jak portale informacyjne, platformy wideo oraz media społecznościowe[98].

W analizie *fake newsów* kluczowe jest odróżnienie pojęć „dezinformacja" i „misinformacja". Dezinformacja jest świadomym i celowym rozpowszechnianiem fałszywych informacji, mającym na celu manipulację odbiorcami. Natomiast misinformacja dotyczy niezamierzonego propagowania błędnych treści, wynikających na przykład z nieświadomego błędu dziennikarskiego[99]. Do najczęściej stosowanych technik manipulacji treścią należą *clickbaitowe* nagłówki, które wykorzystują sensacyjny język oraz odwołują się do silnych emocji, takich jak oburzenie czy lęk, w celu zwiększenia zainteresowania odbiorców[100]. Typowe techniki manipulacji audiowi-

[96] A. Gelfert, Fake news: a definitione, Informal Logic 38, 2018, s. 84-117.

[97] Zob. E. C. Tandoc, Z. W. Lim, R. Ling, Defining fake news. A typology of scholarly definitions, Digital Journalism, 6(2), s. 1-17.

[98] Zob. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science, 359(6380), 1146-115.

[99] Zob. C. Wardle, C. H. Derakhshan, Information Disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe report DGI(2017)09, 2017.

[100] Zob. V. Bakir, A. McStay, Fake News and the Economy of Emotions: Problems, Causes, Solutions, Digital Journalism, 6(2), 154-175.

zualnej, takie jak selektywny montaż, fałszywy kontekst obrazów (*miscaptioning*), manipulacje zdjęciowe i *deepfake'i*, szczegółowo przedstawione zostały w tabeli 1. Tego typu manipulacje mają na celu stworzenie wrażenia prawdziwości i wiarygodności prezentowanych treści[101].

Tabela 1. Typowe techniki tworzenia fake news i manipulacji audiowizualnej

Technika manipulacji	Opis charakterystyczny	Przykład praktyczny
Clickbait i język emocji	Stosowanie przesadzonych, sensacyjnych nagłówków i sformułowań, odwołujących się do ciekawości lub silnych emocji odbiorcy, celem przyciągnięcia uwagi.	Artykuł zatytułowany *„Szok! Nie uwierzysz, co znaleziono…"* sugeruje rewelację, choć treść jest błaha lub fałszywa.
Fałszywy kontekst	Przedstawianie prawdziwego obrazu lub wideo w mylnym kontekście – zmiana opisu, daty lub okoliczności, by nadać materiałowi nowe znaczenie.	Zdjęcie z katastrofy sprzed lat opatrzone bieżącym podpisem informującym o rzekomym wydarzeniu dziś.
Manipulacja montażem	Edycja materiałów wideo/audio poprzez cięcia, łączenie ujęć lub zmiany prędkości, aby zmienić wydźwięk lub interpretację wypowiedzi.	Zmontowany film, w którym wyrwane z kontekstu słowa polityka tworzą wrażenie skandalu, choć w oryginale sens był inny.
Deepfake (głębokie fałszerstwo)	Wykorzystanie algorytmów SI do wygenerowania realistycznego obrazu lub nagrania przedstawiającego osobę robiącą lub mówiącą coś, co nigdy nie miało miejsca.	Fałszywe wideo, na którym znany aktor „przyznaje się" do czynu, którego nie popełnił (wygenerowane na podstawie jego wizerunku).
Trolle i *boty* internetowe	Skoordynowane działania wielu fałszywych kont lub programów automatycznych rozpowszechniających określone treści na masową skalę w mediach społecznościowych.	„Zmasowany atak" dezinformacji na Twitterze: setki nowych kont jednocześnie udostępniają ten sam link do fałszywej wiadomości.

Źródło: opracowanie własne.

Szczególnie istotną rolę w rozprzestrzenianiu *fake newsów* pełnią zorganizowane struktury, zwane „fabrykami trolli", które za pomocą kont internetowych (botów) i zautomatyzowanych systemów rozpowszechniają zmanipulowane informacje, zwielokrotniając ich widoczność i oddziaływanie społeczne. Psychologiczne mechanizmy podatności odbiorców na *fake newsy* są istotnym elementem ich efektywności. Jednym z głównych czynników jest efekt potwierdzenia (ang. *confirmation bias*), polegający na akceptacji informacji zgodnych z wcześniejszymi przekonaniami odbiorców, a odrzucaniu tych, które stoją z nimi w sprzeczności. Dodatkowym mechanizmem sprzyjającym dezinformacji jest stosowanie heurystyk poznawczych, które

[101] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, 1753-1820.

w warunkach przeciążenia informacyjnego prowadzą do uproszczonej oceny wiarygodności źródła i treści. Ważną rolę odgrywają również emocje, gdyż informacje wywołujące silne reakcje emocjonalne, takie jak gniew, strach lub zaskoczenie, sprzyjają szybkiemu i bezrefleksyjnemu rozpowszechnianiu treści. Ponadto algorytmy mediów społecznościowych powodują powstawanie baniek informacyjnych, co dodatkowo wzmacnia mechanizm selektywnej ekspozycji oraz ogranicza konfrontację z odmiennymi opiniami, co ostatecznie zwiększa podatność odbiorców na manipulację.

Fake newsy i techniki manipulacji audiowizualnej stanowią poważne zagrożenie dla bezpieczeństwa informacyjnego państw. Mechanizmy ich tworzenia oraz psychologiczne uwarunkowania podatności odbiorców na manipulacje wymagają interdyscyplinarnych działań zaradczych. Do najważniejszych należą regulacje prawne, edukacja społeczeństwa oraz rozwój technologii umożliwiających szybką identyfikację i neutralizację dezinformacji. Regulacje prawne dotyczące *fake newsów* obejmują m.in. przepisy mające na celu odpowiedzialność autorów i dystrybutorów za świadome rozpowszechnianie dezinformacji. Przykładowo, w Unii Europejskiej, Dyrektywa o usługach cyfrowych (ang. *Digital Services Act* – DSA) nakłada na platformy internetowe obowiązek przeciwdziałania rozpowszechnianiu *fake newsów* oraz zwiększenia transparentności algorytmów odpowiedzialnych za ich dystrybucję. W Polsce kwestie związane z dezinformacją regulowane są m.in. przez przepisy kodeksu karnego dotyczące zniesławienia oraz rozpowszechniania nieprawdziwych informacji mogących wywołać niepokój publiczny. Rozwiązania prawne w skali międzynarodowej wymagają jednak ciągłego doskonalenia i adaptacji do zmieniających się technik manipulacji audiowizualnej i sposobów dystrybucji *fake newsów*, co jest niezbędne, by skutecznie ograniczyć ich negatywny wpływ na współczesne społeczeństwa.

4.3. Deepfake jako narzędzie destabilizacji politycznej i militarnej

Technologia *deepfake*, powstała z połączenia terminów *deep learning* oraz *fake*, polega na tworzeniu realistycznych obrazów i materiałów audio-wizualnych przy użyciu zaawansowanych algorytmów sztucznej

inteligencji[102]. Kluczowym mechanizmem tej koncepcji pozostają generatywne sieci przeciwstawne (ang. *Generative Adversarial Networks* – GAN), składające się z generatora, który tworzy syntetyczne dane w oparciu o zestawy treningowe, oraz dyskryminatora, oceniającego ich wiarygodność[103]. Iteracyjne uczenie się obu modeli umożliwia wytwarzanie materiałów trudnych do odróżnienia od autentycznych. Początkowo technologia deepfake zyskała rozgłos w 2017 roku za sprawą publikacji na platformie Reddit, gdzie użytkownik o pseudonimie „*deepfakes*" zamieścił przerobione filmy z udziałem celebrytów[104]. Wówczas postrzegano ją głównie jako ciekawostkę z pogranicza rozrywki i efektów specjalnych. Z czasem ujawnił się jednak jej potencjał destabilizacyjny, zwłaszcza w sferach politycznej i militarnej, co wpisuje się w szerszy kontekst bezpieczeństwa państwa. Dostępność zaawansowanych narzędzi do generowania fałszywych materiałów audio-wizualnych wzmocniła obawy o ich wykorzystanie do podważania zaufania publicznego i kreowania dezinformacji[105].

Z perspektywy bezpieczeństwa państwa istotne jest uwzględnienie wieloaspektowego wykorzystania *deepfake* w polityce i sferze militarnej. W obszarze politycznym technologia ta może służyć do tworzenia materiałów dyskredytujących przywódców lub przeciwników politycznych poprzez przypisywanie im wypowiedzi czy zachowań, których nigdy nie przejawiali[106]. Takie działania mogą zakłócać procesy demokratyczne, zwłaszcza w czasie kampanii wyborczych, kiedy opinia publiczna jest szczególnie podatna na wpływy dezinformacyjne. Z kolei w wymiarze wojskowym *deepfake* może stanowić komponent operacji psychologicznych, mających na celu obniżenie morale wojsk lub wywołanie zamieszania wśród ludności cywilnej. Sfabrykowane komunikaty przypisywane przywódcom państwo-

[102] Zob. 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), 255-262.

[103] Zob. M. Westerlund, The emergence of deepfake technology: A review, Technology Innovation Management Review, 9(11), 39-52.

[104] Zob. T. Dobber, N. Metoui, D. Trilling, N. Helberger, C. de Vreese, Do (microtargeted) deepfakes have real effects on political attitudes?, Harvard Kennedy School Misinformation Review, 1(Special Issue on Artificial Intelligence and Media Integrity), 2020.

[105] Zob. R. Chesney, D. Citron, Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107(1753), 1753-1820.

[106] Zob. C. Constantineanu, Deepfakes: The New Frontier in Political Disinformation, The Security Distillery, 2023.

wym mogą prowadzić do dezorientacji w czasie konfliktu zbrojnego lub napięć międzynarodowych.

Rozwój technologii *deepfake* nie ogranicza się wyłącznie do państw dysponujących zaawansowanym zapleczem technologicznym. Łatwo dostępne oprogramowanie oraz otwarte zasoby internetowe sprzyjają demokratyzacji narzędzi generatywnych, co zwiększa ryzyko, że grupy ekstremistyczne, przestępcze czy niezależni aktorzy działający z pobudek politycznych wykorzystają je do destabilizacji sytuacji wewnętrznej w państwach. Wielowymiarowe zagrożenia obejmują tym samym nie tylko konfrontacje zewnętrzne, ale również konflikty wewnątrzpaństwowe i próby wywierania wpływu na lokalne społeczności.

Skała zagrożenia płynącego z technologii *deepfake* w znacznej mierze wynika z jej zdolności do realistycznego zniekształcania rzeczywistości, co może ułatwiać podważanie stabilności państw i społeczeństw. Z perspektywy bezpieczeństwa narodowego kluczowe znaczenie mają skutki w wymiarze politycznym: sfabrykowane materiały rozpowszechniane tuż przed wyborami mogą wpływać na preferencje wyborców, a powszechne użycie *deepfake* przyczynia się do erozji zaufania w przestrzeni publicznej. Jednocześnie świadomość istnienia technologii *deepfake* powoduje, że odbiorcy mogą sceptycznie podchodzić nawet do wiarygodnych materiałów, co sprzyja chaosowi informacyjnemu i zjawisku postprawdy. Dodatkowo, podmioty polityczne mogą wykorzystywać tę niepewność, odrzucając autentyczne dowody i tłumacząc je jako próby manipulacji.

Kolejnym poważnym wyzwaniem jest ryzyko przeprowadzania operacji psychologicznych skierowanych zarówno przeciwko cywilom, jak i wojskom przeciwnika. *Deepfake* może służyć do generowania fałszywych przemówień nakłaniających do kapitulacji, tworzenia spreparowanych informacji o rzekomych okrucieństwach lub pokazywania fikcyjnych wydarzeń uzasadniających agresję militarną. W efekcie może dochodzić do wzmacniania niepokojów społecznych oraz eskalacji konfliktów na skalę międzynarodową. Dla służb bezpieczeństwa koniecznością staje się zatem opracowanie skutecznych procedur rozpoznawania i przeciwdziałania tego rodzaju manipulacjom.

Media społecznościowe odgrywają szczególną rolę w procesie dyfuzji deepfake, ponieważ sprzyjają one błyskawicznemu rozprzestrzenianiu treści generujących silne emocje. Zjawisko to niesie poważne implikacje dla

bezpieczeństwa państwa, gdyż w krótkim czasie możliwe staje się dotarcie do szerokiego grona odbiorców. Wirusowy charakter postów w mediach internetowych oraz brak jednolitego i skutecznego systemu weryfikacji materiałów powodują, że potencjalne szkody wywołane przez dezinformację mogą narastać w sposób trudny do kontrolowania[107]. Ponadto zaawansowane algorytmy rekomendujące, zaprojektowane w celu zwiększania zaangażowania użytkowników, często promują kontrowersyjne lub szokujące materiały, co dodatkowo ułatwia szybkie rozprzestrzenianie się sfabrykowanych treści.

Istotnym wyzwaniem jest także ciągła ewolucja technologii. Algorytmy detekcji, choć opracowywane przez instytucje naukowe oraz przedsiębiorstwa technologiczne, często ustępują tempem rozwoju algorytmom generatywnym, co czyni walkę z *deepfake* swego rodzaju „wyścigiem zbrojeń" w środowisku cyfrowym. Z uwagi na potencjał destabilizacyjny, zarówno organy państwowe, jak i prywatne platformy medialne muszą podejmować współpracę mającą na celu nie tylko ograniczanie rozprzestrzeniania nieprawdziwych informacji, ale również edukację społeczeństwa w zakresie krytycznej analizy treści.

W kontekście bezpieczeństwa państwa szczególnym przykładem jest kryzys polityczny w Gabonie w 2019 roku. Podejrzenia dotyczące użycia *deepfake* do sfałszowania przemówienia prezydenta Alego Bongo przyczyniły się do wzrostu napięć i próby wojskowego przewrotu, choć późniejsza ekspertyza nie stwierdziła jednoznacznego fałszerstwa. Incydent ten odzwierciedla skalę zagrożenia, jaką niesie samo podejrzenie manipulacji. Innym przypadkiem wykorzystania deepfake pozostaje konflikt rosyjsko-ukraiński z 2022 roku, gdzie spreparowane nagranie przypisywane prezydentowi Ukrainy Wołodymyrowi Zełenskiemu wzywało do złożenia broni. Szybkie zdementowanie tych treści ograniczyło skalę szkód, jednak zdarzenie potwierdziło użyteczność tej technologii w wojnie informacyjnej. Z kolei podszywanie się pod mera Kijowa Witalija Kliczkę w czerwcu 2022 roku doprowadziło do dezinformowania samorządowców w kilku europejskich metropoliach, unaoczniając rosnące ryzyko w obszarze relacji dyplomatycznych. Tego rodzaju incydenty wskazują, że technologia *deepfake* nie ogranicza się do sfery propagandy wojennej, lecz wpływa także na prowadzenie polityki międzynarodowej i wewnętrznej.

[107] Zob. A. Kaur, A. Dhiman, M. Singh, Comprehensive Review: Security Challenges and Countermeasures for Big Data Security in Cloud Computing, IEEE, 2023.

Dynamiczny rozwój *deepfake* stanowi istotne wyzwanie z perspektywy bezpieczeństwa państwa i stabilności globalnej. Realistyczne fałszerstwa audio-wizualne mogą kreować fikcyjne narracje, które podważają fundamenty zaufania społecznego i destabilizują procesy demokratyczne. Skalę zagrożeń powiększa szeroka dostępność technologii, pozwalająca zarówno aktorom państwowym, jak i niepaństwowym, wykorzystywać *deepfake* do wywierania wpływu na scenę polityczną, prowadzenia operacji psychologicznych czy dezinformacji społeczeństwa. Rola mediów społecznościowych w tym kontekście pozostaje kluczowa, gdyż umożliwiają one masowe rozpowszechnianie fałszywych materiałów w niezwykle krótkim czasie. Skuteczne przeciwdziałanie zagrożeniom generowanym przez *deepfake* wymaga zatem kompleksowego podejścia łączącego działania technologiczne, prawne i edukacyjne. Regularne doskonalenie narzędzi detekcji, odpowiednie regulacje prawne penalizujące szkodliwe użycie *deepfake* oraz kształtowanie świadomości społecznej w zakresie krytycznej analizy treści mogą przyczynić się do ograniczenia negatywnych konsekwencji związanych z manipulacjami medialnymi. W dobie przyspieszonego rozwoju sztucznej inteligencji kluczowe staje się zachowanie równowagi między postępem technologicznym a koniecznością ochrony stabilności społeczno-politycznej i bezpieczeństwa państwa[108].

4.4. Analiza przypadków wpływu dezinformacji audiowizualnej na bezpieczeństwo państwa

We współczesnej erze cyfrowej, charakteryzującej się powszechnym dostępem do zaawansowanych technologii komunikacyjnych, obserwuje się intensyfikację zagrożeń wynikających z dezinformacji audiowizualnej. Termin ten odnosi się do celowo spreparowanych lub zmanipulowanych materiałów w formie nagrań wideo, plików audio oraz obrazów, których celem jest wprowadzenie odbiorców w błąd, a tym samym wywarcie określonego wpływu społeczno-politycznego. Zaufanie, jakim tradycyjnie darzono przekazy wizualne (oparte na założeniu „widzieć to znaczy wierzyć"), uległo erozji

[108] Zob. T. Protasowicki, J. Stanik, Big Data w analizie zagrożeń bezpieczeństwa narodowego, Ekonomiczne Problemy Usług, 123, 2016.

w obliczu możliwości ich realistycznego fałszowania, co umożliwia skuteczne dezinformowanie opinii publicznej oraz destabilizację ładu informacyjnego.

Skala problemu znajduje potwierdzenie w licznych udokumentowanych przypadkach, w których zmanipulowane treści audiowizualne wykorzystywano jako instrumenty wpływu politycznego, narzędzia destabilizacji społecznej, środki ingerencji w procesy demokratyczne – w tym wyborcze – oraz komponenty działań z zakresu cyberwojny i operacji informacyjnych prowadzonych w ramach konfliktów zbrojnych. Przedmiotowy rozdział prezentuje syntetyczny przegląd wybranych studiów przypadków, poddając analizie ich konsekwencje dla bezpieczeństwa narodowego oraz identyfikując strategie przeciwdziałania tym zagrożeniom. Wśród najbardziej znanych incydentów ilustrujących zjawisko dezinformacji audiowizualnej w przestrzeni politycznej znajduje się przypadek zmanipulowanego nagrania z 2019 roku, przedstawiającego przewodniczącą Izby Reprezentantów Stanów Zjednoczonych, Nancy Pelosi[109]. W mediach społecznościowych rozpowszechniano spreparowany materiał wideo, w którym celowo spowolniono tempo jej wypowiedzi, aby wywołać fałszywe wrażenie, iż polityk znajduje się pod wpływem alkoholu lub cierpi na zaburzenia neurologiczne. Mimo że technika manipulacji była relatywnie prymitywna (tzw. *cheapfake*), nagranie osiągnęło duży zasięg i przyczyniło się do eskalacji dezinformacyjnych narracji w debacie publicznej[110]. Choć technika edycyjna zastosowana w omawianym przypadku była relatywnie prymitywna – określana mianem *cheapfake* lub *shallowfake* – materiał zyskał ogromny zasięg w przestrzeni cyfrowej. Co istotne, jego dystrybucję wspierali prominentni przedstawiciele życia politycznego: doradca prezydenta Stanów Zjednoczonych, Rudy Giuliani, a następnie sam prezydent Donald Trump udostępnili wideo w serwisie Twitter, nadając mu pozory autentyczności i podnosząc jego percepcję jako wiarygodnego źródła. Pomimo szybkiej reakcji mediów głównego nurtu, które zdementowały autentyczność nagrania, jego zasięg znacząco przekroczył zasięg sprostowania. Efekt ten ukazuje jedną z kluczowych cech dezinformacji w środowisku cyfrowym: przewagę narracji pierwotnej (nawet fałszywej) nad późniejszym sprostowaniem, znaną w literaturze jako

[109] Zob. C. Funk, The Ethics of Cheapfakes: The Case of Nancy Pelosi, Center for Media Engagement, University of Texas at Austin, 2021.

[110] Zob. R. Chesney, D. Citron, Q. Jurecic, About That Pelosi Video: What to Do About 'Cheapfakes' in 2020, Lawfare, 2019.

efekt pierwszeństwa informacyjnego. Wiele osób zaakceptowało fałszywy przekaz, zwłaszcza że był on spójny z ich wcześniejszymi przekonaniami politycznymi – co potwierdza rolę tzw. efektu potwierdzenia (*confirmation bias*) jako mechanizmu wzmacniającego skuteczność dezinformacji. Opisany incydent unaocznia, iż nawet nieskomplikowane technologicznie manipulacje treściami wideo mogą wywoływać poważne skutki polityczne, przyczyniając się do polaryzacji społecznej, erozji autorytetu liderów oraz pogłębiania nieufności wobec instytucji demokratycznych. Jest to jednoznaczne ostrzeżenie, że do uruchomienia tzw. „kaskady dezinformacyjnej" nie jest konieczne zastosowanie zaawansowanych narzędzi z zakresu sztucznej inteligencji – wystarczy odpowiednie wykorzystanie mechanizmów amplifikacji przekazu charakterystycznych dla mediów społecznościowych oraz poznawcze predyspozycje odbiorców do akceptowania treści zgodnych z ich światopoglądem.

Zjawisko dezinformacji audiowizualnej może również przybierać formę destabilizacji społecznej poprzez stymulowanie paniki moralnej, eskalacji nienawiści lub aktów przemocy motywowanych etnicznie. Wymownym przykładem takiego mechanizmu była fala linczów, która miała miejsce w Indiach w latach 2017–2018, wywołana przez dezinformacyjne przekazy rozpowszechniane za pośrednictwem platform komunikacyjnych[111,112]. Na obszarach wiejskich rozprzestrzeniano drastyczne materiały wideo oraz nagrania głosowe, przedstawiające rzekome akty przemocy przypisywane grupom porywaczy dzieci. Szczególnie rozpowszechnione było nagranie ukazujące brutalne morderstwo, uznane – błędnie – za dowód działalności wędrownego plemienia Parthi. Pomimo że materiał pochodził z innego regionu i był całkowicie sfabrykowany, jego odbiór przez lokalne społeczności wywołał masową panikę, prowadząc do tworzenia oddolnych patroli obywatelskich oraz licznych aktów samosądu. W efekcie, w ciągu kilkunastu miesięcy doszło do co najmniej 69 odnotowanych ataków motłochu, których ofiarą padły 33 osoby, niesłusznie uznane za zagrożenie na podstawie fałszywych doniesień medialnych[113]. Lokalne społeczności, ufając

[111] Zob. R. Bhakt, WhatsApp Rumours and Lynchings: Mob Violence in India, Economic & Political Weekly, 54(6), 2019.

[112] Zob. E. Gowen, How Misinformation on WhatsApp Led to a Deathly Mob Lynching in India, The Washington Post, 2020.

[113] Zob. R. Rajeshwari, Mob Lynching and Social Media, Yale Journal of International Affairs, 14(1), 2019.

informacjom otrzymanym od znajomych i krewnych, dały się wciągnąć w spiralę paniki i przemocy. Dezinformacja rozpowszechniana w mediach społecznościowych przełożyła się bezpośrednio na zagrożenie bezpieczeństwa wewnętrznego – destabilizację porządku publicznego, lincze i wzrost napięć międzygrupowych. Analogiczne mechanizmy propagowania mowy nienawiści oraz dezinformacji zaobserwowano również w kontekście międzynarodowym – przykładowo w Mjanmie rozpowszechnianie fałszywych treści w mediach społecznościowych wymierzonych w mniejszość etniczną Rohingya znacząco przyczyniło się do społecznego przyzwolenia na działania o charakterze czystek etnicznych.[114,115]. Opisane incydenty jednoznacznie potwierdzają, iż dezinformacja audiowizualna może pełnić funkcję katalizatora przemocy zbiorowej, prowadząc do eskalacji napięć społecznych oraz wystąpień o charakterze linczów i samosądów. W kontekście Indii zjawisko to zostało zidentyfikowane przez władze państwowe jako realne zagrożenie dla bezpieczeństwa wewnętrznego, wynikające z niekontrolowanego rozpowszechniania sfabrykowanych treści za pośrednictwem platform społecznościowych. Zintensyfikowana dystrybucja zmanipulowanych materiałów wideo i audio, zawierających fałszywe narracje o charakterze alarmistycznym lub nienawistnym, doprowadziła do sytuacji, w której media społecznościowe zaczęły funkcjonować jako instrumenty masowej paniki informacyjnej. W odpowiedzi na narastające zagrożenie instytucje państwowe wdrożyły szereg działań zaradczych, w tym programy edukacji medialnej oraz bezpośrednie interwencje organów ścigania w społecznościach lokalnych, mające na celu deeskalację napięć oraz wzmacnianie odporności informacyjnej obywateli. Przypadek indyjski dowodzi zatem, że państwo, stojąc wobec gwałtownej dezinformacji wizualnej i dźwiękowej, zmuszone jest niekiedy do uruchomienia nadzwyczajnych środków operacyjnych i prewencyjnych w celu ograniczenia wpływu fałszywych treści oraz przywrócenia poczucia bezpieczeństwa publicznego.

Szczególnie niepokojące są przypadki, w których dezinformacja audiowizualna bezpośrednio ingeruje w procesy demokratyczne, w tym w kampanie wyborcze i mechanizmy legitymizacji władzy politycznej.

---

[114] Amnesty International, Myanmar, We Will Destroy Everything, Military Responsibility for Crimes Against Humanity in Rakhine State, Amnesty International Ltd, London, 2018.

[115] Zob. D. Liebowitz, Social Media and Social Cleansing, Facebook and the Rohingya Crisis, 2018.

Obawy związane z tego rodzaju zjawiskami zaczęły narastać już w drugiej połowie drugiej dekady XXI wieku, kiedy pojawiły się pierwsze przykłady tzw. *deepfake'ów* – zaawansowanych technologicznie, syntetycznie generowanych nagrań audiowizualnych, imitujących wygląd i głos znanych postaci publicznych, w sposób sugerujący, że wypowiadają one treści, których w rzeczywistości nigdy nie wyraziły. Początkowo problem ten miał charakter hipotetyczny i był analizowany głównie w ramach studiów nad potencjalnymi zagrożeniami płynącymi z rozwoju sztucznej inteligencji. Jednak w kolejnych latach zaczęły pojawiać się realne przykłady wykorzystania tego rodzaju technologii w celach politycznych. Przykład stanowi incydent z kampanii przed wyborami parlamentarnymi na Słowacji w 2023 roku, kiedy to na dwa dni przed głosowaniem w mediach społecznościowych opublikowano rzekome nagranie audio, w którym lider proeuropejskiej opozycji, Michal Šimečka, miał rozmawiać z dziennikarzem o planowanym fałszowaniu wyników wyborów. Zarówno polityk, jak i domniemany rozmówca niezwłocznie zdementowali autentyczność nagrania, wskazując na jego fałszywy charakter i możliwe generatywne pochodzenie. Niemniej jednak materiał ten zdążył w krótkim czasie osiągnąć zasięg *wiralowy*, trafiając do szerokiej grupy odbiorców w kluczowym momencie kampanii wyborczej – tuż przed rozpoczęciem ciszy wyborczej. Incydent ten stanowi wyraźne potwierdzenie, że technologia *deepfake* może zostać wykorzystana do celowej manipulacji opinią publiczną, naruszenia integralności procesów wyborczych oraz podważania zaufania społecznego do demokratycznych instytucji.[116]. Wielu komentatorów politycznych i analityków mediów zwracało uwagę, że wspomniany „przeciek" ze Słowacji – zrealizowany najprawdopodobniej przy wykorzystaniu technologii *deepfake audio* – mógł realnie wpłynąć na preferencje elektoratu, odbierając kluczowe głosy proeuropejskiej opozycji i przyczyniając się tym samym do zwycięstwa ugrupowania prorosyjskiego, które oparło swoją kampanię na strategiach dezinformacyjnych. Niezależnie od trudności w jednoznacznym oszacowaniu wpływu tego zdarzenia na końcowy rezultat wyborów, tzw. „słowacki *deepfake*" unaocznił potencjał technologii sztucznej inteligencji do bezprecedensowej ingerencji w przebieg procesów demokratycznych – zwłaszcza w kontekście społeczeństw o niskim poziomie zaufania instytucjonalnego i wysokiej podatności

[116] L. de Nadal, P. Jančárik, Beyond the deepfake hype: AI, democracy, and "the Slovak case", HKS Misinformation Review, 2024.

na propagandę cyfrową. Podobna sytuacja miała miejsce w Turcji, tuż przed wyborami prezydenckimi w 2023 roku. Kandydat opozycji, Muharrem İnce, zrezygnował z udziału w wyścigu wyborczym zaledwie kilka dni przed planowanym głosowaniem, deklarując, że stał się celem „brudnej kampanii" opartej na zmanipulowanym materiale kompromitującym[117]. Rozpowszechniane w sieci nagranie, rzekomo ukazujące go w sytuacji intymnej, zostało przez samego İnce określone jako *deepfake* stworzony z wykorzystaniem fragmentów materiałów pornograficznych. Kandydat publicznie nazwał tę akcję „oszczerczą próbą zamachu na jego wiarygodność polityczną". Niezależnie od prawdziwości tych twierdzeń, efekt społeczno-polityczny był niezaprzeczalny – wycofanie się z kampanii jednego z głównych konkurentów znacząco przeformatowało układ sił politycznych. Z kolei w tej samej kampanii wyborczej urzędujący prezydent Recep Tayyip Erdoğan wykorzystał zmanipulowane nagranie wideo, które miało powiązać przedstawicieli opozycji z organizacjami terrorystycznymi[118]. Tego rodzaju praktyki obrazują, w jakim stopniu technologia generatywna może zostać zaadaptowana do realizacji celów politycznych, obejmujących zarówno dyskredytację oponentów poprzez fałszywe skandale, jak i kreowanie narracji o rzekomych spiskach czy nadużyciach wyborczych. Dezinformacja audiowizualna coraz częściej staje się narzędziem podważania społecznego zaufania do demokratycznych instytucji oraz procesu wyborczego jako takiego.

Na szczególną uwagę zasługuje również rosnąca liczba przypadków wykorzystania technologii *deepfake* w kontekście cyberprzestępczości i zagrożeń dla bezpieczeństwa ekonomicznego oraz infrastrukturalnego państw. Przykładowo, w 2019 roku dyrektor finansowy brytyjskiej spółki energetycznej padł ofiarą wyrafinowanego ataku socjotechnicznego. Cyberprzestępcy, wykorzystując syntetycznie wygenerowany głos naśladujący prezesa spółki-matki, nakłonili go telefonicznie do przelania 243 tysięcy dolarów na wskazane konto. Incydent ten stał się jednym z pierwszych odnotowanych przypadków tzw. *voice spoofing* opartego na algorytmach sztucznej inteligencji, dowodząc tym samym, że manipulacje audiowizualne mogą być również skutecznie wykorzystywane do ataków na sektor prywatny

[117] Zob. R. Michaelson, Turkish presidential candidate quits race after release of alleged sex tape. The Guardian, 2023.

[118] Recorded Future (Insikt Group, Targets, Objectives, and Emerging Tactics of Political Deepfakes), Research Report, 2024.

i finansowy[119]. Zgodnie z danymi ujawnionymi przez firmę Symantec, już w 2019 roku odnotowano co najmniej trzy przypadki zastosowania technologii *deepfake audio* w ramach ataków socjotechnicznych, w których syntetycznie generowany głos – imitujący tożsamość przełożonego lub osoby z wyższej hierarchii korporacyjnej – posłużył do wyłudzenia znacznych kwot pieniędzy od przedstawicieli kadry zarządzającej. Łączne straty poniesione przez ofiary tych incydentów sięgnęły milionów dolarów[120].

Ataki socjotechniczne z wykorzystaniem syntetycznie generowanych głosów, określane w literaturze jako *deepfake phishing* bądź *audio spoofing*, stają się coraz bardziej powszechnym narzędziem stosowanym przez zorganizowane grupy przestępcze. Rozwój technologii sztucznej inteligencji, w szczególności algorytmów uczenia głębokiego, umożliwia precyzyjne modelowanie cech głosu konkretnej osoby, co w praktyce prowadzi do skutecznego obchodzenia konwencjonalnych metod weryfikacji tożsamości – takich jak rozpoznanie głosu w trakcie rozmowy telefonicznej czy werbalne potwierdzenie dyspozycji służbowej. Z perspektywy bezpieczeństwa instytucjonalnego – zarówno w sektorze publicznym, jak i prywatnym – konsekwencje tego typu działań są szczególnie niebezpieczne. W przypadku skutecznego podszycia się pod głos osoby zajmującej wysokie stanowisko administracyjne, dowódcze bądź zarządcze, możliwe jest wydanie fałszywego polecenia o charakterze operacyjnym. Może to prowadzić do dezintegracji mechanizmów decyzyjnych, zakłócenia działania infrastruktury krytycznej, a w skrajnych przypadkach – stanowić zagrożenie dla życia i zdrowia obywateli. Symbolicznym przykładem zagrożeń wynikających z braku skutecznych procedur weryfikacji tożsamości cyfrowej pozostaje incydent z czerwca 2022 roku, który wywołał zaniepokojenie w środowiskach administracji publicznej państw Unii Europejskiej. Kilku burmistrzów europejskich metropolii – w tym Berlina i Wiednia – odbyło zdalne rozmowy wideo z osobą podszywającą się pod mera Kijowa, Witalija Kłyczkę[121]. Rozmówca

[119] Zob. C. Stupp, Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Wall Street Journal, 2019.

[120] Zob. N. Naik, P. Kumar, R. Aggarwal and S. Upadhyay, AI-Generated Voice Deepfakes and Security Implications: A Systematic Review, Journal of Cybersecurity and Privacy, 2023, s. 115-139.

[121] Zob. J. Twomey, D. Ching, M P. Aylett, M. Quayle., C. Linehan, G. Murphy, Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. PLOS ONE, 18(10), e0289937.

posługiwał się na tyle wiarygodnie odwzorowaną mimiką oraz głosem, że uczestnicy spotkania zorientowali się dopiero po czasie, kiedy pojawiły się nielogiczne i prowokacyjne pytania. Jak później ustalono, za operacją stali rosyjscy prowokatorzy medialni, wykorzystujący najprawdopodobniej technologię deepfake lub inne techniki cyfrowej manipulacji obrazu i dźwięku. Mimo pozornie humorystycznego charakteru incydentu, eksperci ds. cyberbezpieczeństwa wskazali na jego poważne implikacje. Zdarzenie to obnażyło systemowe luki w protokołach identyfikacyjnych wykorzystywanych przez administrację publiczną w warunkach komunikacji zdalnej. Gdyby podobne działanie miało charakter operacji wywiadowczej, mogłoby prowadzić do ujawnienia poufnych informacji lub nieświadomego zaangażowania przedstawicieli władz w działania destabilizujące. Z tego względu dynamiczne rozprzestrzenianie się technologii *deepfake* postrzegane jest jako jedno z kluczowych wyzwań dla bezpieczeństwa informacyjnego i teleinformatycznego w XXI wieku, wymagające zarówno aktualizacji procedur, jak i wzmożonego nadzoru instytucjonalnego.

Najpoważniejsze zagrożenie dla bezpieczeństwa państwa pojawia się wówczas, gdy dezinformacja audiowizualna zostaje włączona do arsenału działań wojny informacyjnej prowadzonej w czasie rzeczywistych konfliktów zbrojnych lub operacji hybrydowych. Choć operacje dezinformacyjne nie są zjawiskiem nowym, dopiero w ostatnich latach zanotowano pierwsze przypadki użycia technologii *deepfake* jako broni psychologicznej na faktycznym polu walki. Przełomowym przykładem była inwazja Federacji Rosyjskiej na Ukrainę w 2022 roku. Już w początkowej fazie konfliktu, w połowie marca, doszło do incydentu polegającego na zhakowaniu ukraińskiego portalu informacyjnego oraz rozpowszechnieniu fałszywego nagrania wideo z udziałem prezydenta Wołodymyra Zełenskiego. W sfabrykowanym materiale wygenerowanym przy użyciu technologii *deepfake* prezydent miał rzekomo wzywać ukraińskie siły zbrojne do złożenia broni i poddania się rosyjskim oddziałom. Równolegle zmanipulowany przekaz pojawił się również na paskach informacyjnych państwowej telewizji Ukraina 24. Choć jakość techniczna materiału pozwoliła na jego stosunkowo szybką dekonspirację, a sam prezydent Zełenski niezwłocznie opublikował autentyczne oświadczenie wideo, które zdementowało fałszywy przekaz, incydent ten stanowił pierwszy udokumentowany przypadek zastosowania *deepfake'u* w kontekście aktywnego konfliktu zbrojnego. Pokazał tym samym potencjał tej

technologii do prowadzenia operacji psychologicznych w czasie rzeczywistym, których celem jest dezinformacja żołnierzy, społeczeństwa i wspólnoty międzynarodowej[122]. W połowie marca 2022 roku, podczas trwającej inwazji Rosji na Ukrainę, doszło także do incydentu związanego z wykorzystaniem technologii *deepfake* w operacjach dezinformacyjnych. Na pasku informacyjnym stacji telewizyjnej Ukraina 24 pojawił się komunikat informujący o rzekomej kapitulacji Ukrainy. Równocześnie rozpowszechniono w mediach społecznościowych sfabrykowane nagranie wideo, w którym prezydent Wołodymyr Zełenski miał wzywać ukraińskie siły zbrojne do złożenia broni. Analiza materiału wykazała, że zarówno obraz, jak i głos prezydenta zostały sztucznie wygenerowane przy użyciu technologii *deepfake*. Jakość techniczna tego fałszywego nagrania była niska, co pozwoliło wielu odbiorcom na szybkie zidentyfikowanie manipulacji. Sam prezydent Zełenski niezwłocznie opublikował autentyczne oświadczenie wideo, w którym zdementował fałszywe informacje i podkreślił determinację Ukrainy w obronie suwerenności. Niemniej jednak, incydent ten wzbudził szerokie zainteresowanie mediów i opinii publicznej, stanowiąc pierwszy udokumentowany przypadek użycia technologii *deepfake* w konflikcie zbrojnym w celu manipulacji masowym odbiorcą oraz wpłynięcia na morale zarówno ludności cywilnej, jak i personelu wojskowego. Chociaż w tym przypadku próba dezinformacji nie osiągnęła zamierzonego efektu i nie podważyła determinacji ukraińskich sił zbrojnych, eksperci do spraw bezpieczeństwa informacyjnego zwrócili uwagę na potencjalne ryzyko związane z bardziej zaawansowanymi technicznie manipulacjami. Gdyby jakość fałszywego nagrania była wyższa, a jego dystrybucja bardziej skoordynowana, skutki mogłyby być znacznie poważniejsze. Incydent ten unaocznił nowy wymiar zagrożeń hybrydowych, w których przeciwnik wykorzystuje zaawansowane technologie do prowadzenia operacji dezinformacyjnych w czasie rzeczywistym, równolegle z działaniami militarnymi. Dodatkowo, przypadek ten uwydatnił zjawisko określane jako „dywidenda kłamcy" (ang. *liar's dividend*). Termin ten odnosi się do sytuacji, w której rosnąca świadomość istnienia technologii umożliwiających tworzenie realistycznych fałszerstw audiowizualnych prowadzi do podważania wiarygodności autentycznych materiałów.

[122] J. Twomey, D. Ching, M P. Aylett, M. Quayle, M., C. Linehan, G. Murphy, G.,Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. PLOS ONE, 18(10), e0289937.

W kontekście konfliktu na Ukrainie zaobserwowano przypadki, w których rzeczywiste nagrania były fałszywie oskarżane o bycie *deepfake'ami*, co prowadziło do dezinformacji i zwiększonej nieufności wśród odbiorców. Zjawisko to ma istotne implikacje dla bezpieczeństwa informacyjnego, gdyż sama świadomość możliwości istnienia *deepfake'ów* może prowadzić do podważania autentyczności prawdziwych komunikatów, co w konsekwencji utrudnia odróżnienie faktów od fałszu w przestrzeni publicznej[123].

Przytoczone przypadki stanowią wyraźną ilustrację wielowymiarowego oddziaływania dezinformacji audiowizualnej na bezpieczeństwo narodowe. Jej skutki obejmują zarówno erozję porządku konstytucyjnego i destabilizację ładu publicznego, jak również potencjalne obniżenie zdolności obronnych państwa w warunkach zagrożenia zewnętrznego. W świetle narastającej skali i złożoności tego typu zagrożeń, niezbędne staje się sformułowanie adekwatnych strategii przeciwdziałania, opartych na zintegrowanym i interdyscyplinarnym podejściu. Doświadczenia ostatnich lat wskazują, iż skuteczna odpowiedź na zjawisko dezinformacji audiowizualnej wymaga zastosowania kompleksowych rozwiązań obejmujących działania w czterech kluczowych obszarach: technologicznym, edukacyjnym, regulacyjnym oraz międzynarodowym. Coraz częściej w dokumentach strategicznych oraz raportach instytucji odpowiedzialnych za bezpieczeństwo narodowe podkreślana jest konieczność koordynacji tych komponentów w ramach całościowej polityki ochrony informacyjnej. W zakresie technologii, rządy oraz instytucje publiczne – w tym Ministerstwo Obrony Narodowej Rzeczypospolitej Polskiej czy amerykańska agencja *Defense Advanced Research Projects Agency* (DARPA) – aktywnie wspierają rozwój zaawansowanych systemów służących detekcji syntetycznie wygenerowanych treści, w szczególności tzw. *deepfake'ów*. Narzędzia te opierają się na algorytmach analizy obrazu i dźwięku, które umożliwiają identyfikację charakterystycznych artefaktów cyfrowych, anomalii temporalnych oraz niespójności między modalnościami przekazu (np. obrazem i dźwiękiem). Ich celem jest skuteczne rozpoznawanie manipulacji audiowizualnych, które mogłyby zostać wykorzystane do destabilizacji procesów politycznych, społecznych lub wojskowych[124].

[123] J. Delcker, Welcome to the age of uncertainty. Deepfake videos may have destabilized national politics for the first time, Politico Europe, 2019.

[124] Zob. Ministerstwo Obrony Narodowej RP, NATO DIANA – Druga edycja konkursu Akcelerator Innowacji Obronnych, Ministerstwo Obrony Narodowej RP, 2021.

Niemniej, jak zauważają eksperci Narodowego Centrum Cyberbezpieczeństwa Wielkiej Brytanii[125], nawet najnowocześniejsze rozwiązania technologiczne nie zapewniają pełnej skuteczności, gdyż twórcy fałszerstw nieustannie doskonalą metody obchodzenia zabezpieczeń. Istotnym elementem strategii przeciwdziałania zagrożeniom dezinformacji audiowizualnej jest edukacja społeczna. Rządowe raporty wskazują na konieczność rozwijania kompetencji cyfrowych obywateli, szczególnie umiejętności krytycznej analizy treści medialnych[126]. Przykładowo, rząd Indii skutecznie ograniczył skalę przemocy wywołanej dezinformacją audiowizualną, realizując kampanie informacyjne obejmujące zarówno nowoczesne środki komunikacji, jak i tradycyjne metody docierania do społeczności wiejskich[127]. Jednakże rząd Wielkiej Brytanii przestrzega przed nadmiernym akcentowaniem zagrożeń *deepfake'ami*, wskazując, że może to paradoksalnie zwiększyć nieufność społeczeństwa wobec wszelkich przekazów medialnych – zjawisko znane jako „dywidenda kłamcy". Z tego względu promuje się strategię *pre-bunkingu,* czyli uprzedzania obywateli o potencjalnych manipulacjach w sposób wyważony, aby nie wywoływać powszechnej podejrzliwości. Kolejnym kluczowym komponentem strategii przeciwdziałania zagrożeniom wynikającym z *deepfake'ów* są regulacje prawne. W Stanach Zjednoczonych wprowadzono przepisy penalizujące rozpowszechnianie fałszywych materiałów audiowizualnych, szczególnie podczas kampanii wyborczych[128]. Podobne inicjatywy podjęła Unia Europejska, rozważając włączenie odpowiednich zapisów dotyczących *deepfake'ów* do planowanego Aktu o sztucznej inteligencji[129]. Platformy internetowe, jak Facebook czy Twitter, również zaktualizowały swoje polityki dotyczące syntetycznych mediów, zobowiązując się do oznaczania lub usuwania zmanipulowanych[130]. Ponadto, skuteczność

[125] Zob. NCSC UK, Disinformation and deepfakes, How to defend against misinformation, National Cyber Security Centre, UK Government, 2022.

[126] NATO StratCom, About NATO Strategic Communications Centre of Excellence, NATO StratCom Centre of Excellence, 2022.

[127] Ministry of Home Affairs, Government of India, Advisory on Fake News and Rumors, Government of India, 2022.

[128] U.S. Congress, Identifying Outputs of Generative Adversarial Networks Act, 116th Congress, Public Law 116-258, 2020.

[129] European Commission, Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), European Commission, 2021.

[130] Meta Transparency Center, Community Standards – Manipulated Media Policy, Meta Platforms Inc. 2022.

działań na poziomie krajowym jest coraz częściej uzupełniana o współpracę międzynarodową i międzysektorową. Organizacja Traktatu Północnoatlantyckiego NATO[131] powołało specjalistyczne centrum przeciwdziałania dezinformacji – NATO StratCom Centre of Excellence, które koordynuje działania sojuszników oraz dostarcza analiz zagrożeń informacyjnych[132]. Z kolei Komisja Europejska uruchomiła platformę *Rapid Alert System*, umożliwiającą szybką wymianę informacji między państwami członkowskimi na temat kampanii dezinformacyjnych, w tym wykorzystujących media audiowizualne[133]. W celu zwiększenia zaufania do autentycznych materiałów audiowizualnych, instytucje państwowe wspierają także rozwój standardów technologicznych, takich jak kryptograficzne znakowanie nagrań na etapie ich powstawania. Rozwiązania te mają zagwarantować integralność danych i pozwolić na łatwiejszą identyfikację manipulacji[134].

Podsumowując, rządowe i ministerialne raporty zgodnie wskazują, że dezinformacja audiowizualna stanowi istotne, wielowymiarowe zagrożenie dla bezpieczeństwa narodowego. Państwa muszą więc podejmować działania interdyscyplinarne, obejmujące technologie obronne, edukację społeczną, regulacje prawne oraz międzynarodową współpracę, aby skutecznie przeciwdziałać skutkom manipulacji audiowizualnych i chronić stabilność społeczną oraz polityczną.

4.5. Strategie przeciwdziałania dezinformacji w różnych państwach

Rosnące zagrożenia ze strony *fake newsów* i *deepfake'ów* skłoniły państwa oraz instytucje międzynarodowe do podjęcia działań zaradczych.

---

[131] NATO (ang. North Atlantic Treaty Organization), Organizacja Traktatu Północnoatlantyckiego, założona w 1949 roku, jest międzynarodowym sojuszem wojskowym zrzeszającym państwa Europy Północnej, Ameryki Północnej i innych regionów, którego celem jest zapewnienie wspólnej obrony oraz utrzymanie pokoju i bezpieczeństwa na arenie międzynarodowej.

[132] NATO StratCom, NATO Strategic Communications Centre of Excellence – Mission and Tasks, NATO, 2022.

[133] European External Action Service, Rapid Alert System: Strengthening coordinated response to disinformation, European Union, 2022.

[134] NIST, Security Measures for 'Deepfake' and AI-Generated Media: A Guide, National Institute of Standards and Technology, 2021.

Strategie te można podzielić na kilka głównych obszarów: regulacje prawne, samoregulacje branżowe, rozwój technologii detekcji oraz edukację i uświadamianie społeczeństwa. Poniżej przedstawiono wybrane inicjatywy – z uwzględnieniem rozwiązań w Polsce i innych krajach – mające na celu zwiększenie odporności państw na dezinformację. Niektóre państwa zdecydowały się na wprowadzenie odrębnych ustaw wymierzonych w zjawisko *fake newsów*, zwłaszcza w kontekście ochrony procesów wyborczych. Przykładem jest Francja, która pod koniec 2018 r. uchwaliła specjalną ustawę przeciw manipulacji informacją w okresie kampanii wyborczych[135]. Prawo to nakłada na wielkie platformy internetowe obowiązek zapewniania wysokiej przejrzystości – w trakcie kampanii muszą one ujawniać, kto finansuje materiały wyborcze i o jakim charakterze, a także mogą zostać zobligowane przez sąd do szybkiego usunięcia rozpowszechnianych *fake newsów* zagrażających uczciwości wyborów. Ustawa francuska wprowadziła również procedury umożliwiające kandydatom lub partiom poszkodowanym przez dezinformację w trybie przyspieszonym dochodzić sprostowań lub blokady szkodliwych treści. Innym modelem jest podejście Niemiec, które wcześniej, bo w 2017 r., przyjęły ustawę *Netzwerkdurchsetzungsgesetz* (NetzDG) dotyczącą walki z bezprawnymi treściami w sieciach społecznościowych[136]. Choć koncentruje się ona na mowie nienawiści i przestępstwach zniesławienia, obejmuje także jawnie fałszywe informacje. Ustawa zobowiązuje duże platformy (ponad 2 mln użytkowników w Niemczech) do sprawnego reagowania na zgłoszenia: nielegalne treści muszą zostać usunięte w ciągu 24 godzin, a pozostałe szkodliwe treści – w ciągu 7 dni. Wśród tych treści wymieniono m.in. fałszywe wiadomości godzące w dobra chronione prawem. Niedostosowanie się do tych wymogów grozi surowymi karami finansowymi, zaś platformom notorycznie łamiącym przepisy NetzDG grożą grzywny do 50 mln euro. Choć prawo to budziło kontrowersje (zarzut ograniczania wolności słowa poprzez prywatną cenzurę), stanowiło pionierski krok wymuszający na korporacjach większą odpowiedzialność za treści

[135] Assemblée nationale (Zgromadzenie Narodowe Francji), Proposition de loi relative à la lutte contre la manipulation de l'information, 15. kadencja, zwyczajna sesja 2018–2019, tekst przyjęty nr 190, 20 listopada 2018.

[136] Bundesrepublik Deutschland, Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), uchwalone 1 września 2017 r. (BGBl. I S. 3352), z późn. zm., ostatnia nowelizacja: art. 29 ustawy z dnia 6 maja 2024, (BGBl. 2024 I Nr. 149).

użytkowników. Inne kraje również poszukiwały rozwiązań legislacyjnych, choć o różnym stopniu restrykcyjności. Wielka Brytania, po serii dochodzeń parlamentarnych (m.in. w sprawie wpływu dezinformacji na referendum *brexit*, zdecydowała się na podejście systemowe. Zamiast odrębnej ustawy o *fake news*, opracowano w 2019 roku Białą Księgę w sprawie szkodliwych treści online[137], która zapoczątkowała prace nad kompleksowymi regulacjami dotyczącymi bezpieczeństwa w Internecie. W efekcie powstał w 2023 roku projekt ustawy o usługach internetowych *Online Safety*[138], nakładającej na platformy obowiązek dbałości o tzw. treści szkodliwe (w tym dezinformację), przy jednoczesnym poszanowaniu wolności słowa. Proponowane przepisy przewidują kodeksy postępowania dla mediów społecznościowych, które mają z jednej strony zwiększyć dostęp użytkowników do rzetelnych informacji, z drugiej – chronić ich przed szkodliwymi przekazami. Co istotne, brytyjskie podejście unika definiowania, co jest prawdą, a co fałszem – ciężar położono na transparentność źródeł informacji i mechanizmów ich dystrybucji (użytkownik ma mieć możliwość sprawdzenia, kto stoi za daną treścią i dlaczego jest mu ona wyświetlana. Takie podejście ma zapobiec nadużyciom i zarazem utrzymać zgodność z zasadą wolności wypowiedzi. W Polsce nie zdecydowano się dotąd na odrębną ustawę skupioną wyłącznie na *fake newsach*. Problematyka ta jest rozproszona w ramach istniejących przepisów – m.in. Prawa prasowego z 1984 r., które zobowiązuje media do prawdziwego przedstawiania zjawisk i dochowywania szczególnej staranności przez dziennikarzy[139]. Publikowanie nieprawdziwych informacji jako takie nie jest przestępstwem (brak jest przepisu karnego sankcjonującego rozpowszechnianie dezinformacji wprost, natomiast autor fałszywej publikacji może ponieść odpowiedzialność cywilną za naruszenie dóbr osobistych lub przepisów branżowych. W polskim systemie prawnym brakuje jednorodnej regulacji dotyczącej dezinformacji – kwestie z tym związane są rozproszone w wielu aktach prawnych różnych dziedzin. Niemniej, w ostatnich latach włączono problem *fake newsów* do strategii bezpieczeństwa

[137] Department for Digital, Culture, Media & Sport and Home Office, Online Harms White Paper, London: HM Government, 2019.

[138] United Kingdom, Online Safety Act 2023 (2023 c.50), London: The Stationery Office, 2023.

[139] KRRiT, Fake news – dezinformacja online. Próby przeciwdziałania tym zjawiskom z perspektywy instytucji międzynarodowych oraz wybranych państw UE, w tym Polski, Warszawa, 2020.

informacji. Na poziomie rządowym koordynacją przeciwdziałania dezinformacji zajmuje się m.in. Rządowe Centrum Bezpieczeństwa (RCB)[140] oraz instytucje związane z cyberbezpieczeństwem (np. NASK, CERT Polska). W 2018 r. uruchomiono portal Bezpieczne Wybory (przed wyborami samorządowymi) jako platformę informującą o zagrożeniach manipulacjami podczas głosowań[141]. Biuro Bezpieczeństwa Narodowego (BBN) uwzględniło dezinformację w analizach zagrożeń hybrydowych[142], a Ministerstwo Cyfryzacji zapowiadało prace nad przepisami weryfikującymi treści w Internecie – choć na poziomie projektów legislacyjnych nie przełożyło się to dotąd na uchwalenie nowego prawa.

Poza twardym prawem, ważną rolę odgrywają samoregulacje i inicjatywy branżowe, w tym działania samych platform oraz porozumienia międzynarodowe o charakterze *soft-law*. W 2018 r. Komisja Europejska zainicjowała Kodeks postępowania w zakresie zwalczania dezinformacji[143], do którego dobrowolnie przystąpiły największe firmy internetowe (m.in. Facebook, Google, Twitter). Zobowiązały się one do m.in. demonetyzacji stron szerzących *fake news* (poprzez odcinanie ich od reklam), promowania wiarygodnych źródeł informacji oraz współpracy z niezależnymi weryfikatorami faktów. Kodeks ten jest narzędziem samoregulacji – nie przewidziano sankcji, ale coroczne raporty monitorują postępy. W 2022 r. zaktualizowano jego zapisy[144], biorąc pod uwagę doświadczenia pandemii (fala dezinformacji covidowej) oraz wojny na Ukrainie. Również tradycyjne media podejmują samoregulacje – np. europejskie stowarzyszenia nadawców przyjęły wytyczne co do przeciwdziałania fałszywym informacjom, a wiele redakcji powołało własne *zespoły fact-checkingowe.* Włochy zdecydowały się na rozwiązanie polegające na współpracy między rządem a platformami i mediami. Włoski Urząd ds. Gwarancji w Komunikacji (*Autorità per le Garanzie nelle Comunicazioni*, AGCOM) w 2017 r. zainicjował forum kooperacyjne (tzw.

[140] Zob. Rządowe Centrum Bezpieczeństwa, Dezinformacja – rosyjska broń strategiczna, Biuletyn, 2021.

[141] NASK, Bezpieczne wybory. Badanie opinii o (dez)informacji w sieci, Warszawa: NASK, 2019.

[142] Biuro Bezpieczeństwa Narodowego, Strategia Bezpieczeństwa Narodowego Rzeczypospolitej Polskiej 2020, Warszawa: BBN, 2020.

[143] Komisja Europejska, Code of Practice on Disinformation, Bruksela: Komisja Europejska, 2018.

[144] Komisja Europejska, Strengthened Code of Practice on Disinformation 2022, Bruksela: Komisja Europejska, 2022.

"okrągły stół techniczny") z udziałem przedstawicieli platform cyfrowych, wydawców prasy, nadawców i dostawców treści[145]. Wypracowano wspólne kodeksy dobrych praktyk – m.in. zalecenia, by platformy podczas kampanii wyborczych udostępniały użytkownikom narzędzia do rozpoznawania *fake newsów* oraz by informowały o mechanizmach funkcjonowania algorytmów rozpowszechniania treści. Ponadto AGCOM zlecił badania nad wpływem platform na system informacyjny oraz nad łańcuchem wartości internetowej dezinformacji, których wyniki posłużyły do sformułowania rekomendacji legislacyjnych[146]. Norwegia z kolei postawiła na wzmocnienie społeczności obywatelskiej i mediów w walce z dezinformacją. Przed wyborami w 2017 r. największe norweskie media powołały wspólnie niezależny portal Faktisk.no, który zajmuje się weryfikacją wypowiedzi polityków, plotek krążących w Internecie i medialnych doniesień. Projekt ten, finansowany przez konsorcjum mediowe i fundację Fritt Ord, stał się w Norwegii jednym z głównych źródeł *fact-checkingu*, a jego opracowania są szeroko cytowane i udostępniane (portal Faktisk szybko stał się jedną z najpopularniejszych stron w kraju). Norwegowie wdrożyli też innowacyjne narzędzia angażujące odbiorców – np. system *Reader Critic* w dzienniku *Dagbladet*, umożliwiający czytelnikom zgłaszanie zauważonych nieścisłości w artykułach online bezpośrednio do redakcji. Takie mechanizmy samokontroli i współpracy z użytkownikami poprawiają jakość informacji u źródła. Nowe technologie wykrywania fałszerstw. Walka z manipulacją audiowizualną to także wyścig technologiczny. Unia Europejska od lat wspiera badania nad narzędziami do automatycznej detekcji i weryfikacji treści online. W programie Horyzont 2020[147] sfinansowano szereg projektów naukowych, m.in.: *InVid* – poświęcony opracowaniu narzędzi do weryfikacji autentyczności zdjęć i nagrań wideo krążących w sieci, Fandango – mający dostarczyć tradycyjnym mediom zaawansowane systemy do wykrywania fałszywych narracji w Internecie, czy *SocialTruth* – tworzący rozproszony ekosystem

[145] AGCOM (Autorità per le Garanzie nelle Comunicazioni), News vs. Fake in the Information System,. Department of Economics and Statistics Report, Rome: AGCOM, November 2018.

[146] AGCOM (Autorità per le Garanzie nelle Comunicazioni), News vs. Fake in the Information System,. Department of Economics and Statistics Report, Rome: AGCOM, November 2018.

[147] European Commission, Horizon 2020 Projects: InVID, Fandango, SocialTruth (CORDIS project database), Brussels: EU, 2018-2021.

ułatwiający dostęp do wielu usług sprawdzania faktów jednocześnie. Rozwijane są także rozwiązania oparte na *blockchainie*, pozwalające śledzić źródło i historię edycji materiałów (tzw. *proof-of-provenance*), co utrudni anonimowe rozpowszechnianie przerobionych mediów. Giganci technologiczni wdrażają własne systemy: YouTube i Facebook używają algorytmów wykrywających znane fejki (np. porównując pliki video z bazami próbek), a Microsoft opracował prototyp aplikacji Video Authenticator, która analizuje stopklatki filmów pod kątem subtelnych artefaktów mogących świadczyć o podmianie twarzy lub głosu. Jednak nawet najlepsze narzędzia na nic się zdadzą, jeśli brakuje współpracy międzynarodowej – stąd inicjatywy takie jak Międzynarodowa Sieć Weryfikacji Faktów (IFCN) przy Instytucie Poyntera, zrzeszająca kilkadziesiąt organizacji *fact-checkingowych* na świecie spełniających wspólne standardy bezstronności i transparentności . Polska również korzysta z tych technologii: zespół CERT Polska (działający w NASK) monitoruje incydenty naruszające bezpieczeństwo sieci i treści, a w razie wykrycia fałszerstw (np. ataków na strony www w celu publikacji *fake news*) nie tylko reaguje doraźnie, ale też rekomenduje instytucjom wdrażanie zabezpieczeń technicznych zapobiegających powtórzeniu się podobnych ataków[148]. Postęp w dziedzinie sztucznej inteligencji działa więc dwukierunkowo: stanowi zagrożenie (*deepfake*), ale i dostarcza broni obronnej (algorytmy detekcji). Długofalowo najważniejszym elementem budowania odporności na dezinformację jest podniesienie kompetencji odbiorców poprzez edukację medialną i świadomość społeczą. Wiele państw integruje edukację medialną z systemem szkolnym i kampaniami publicznymi[149]. Francja równolegle z ustawą anty-dezinformacyjną znowelizowała przepisy oświatowe, wzmacniając nauczanie krytycznego odbioru mediów w szkołach. Uczniowie uczą się odróżniać zweryfikowane źródła od plotek, analizować przekazy wizualne i rozpoznawać typowe techniki manipulacji. Finlandia – często podawana za wzór – już po 2014 r. (agresji rosyjskiej na Krym) wprowadziła powszechne programy edukacyjne uczące

[148] CERT Polska, Raport roczny 2020: Krajobraz bezpieczeństwa polskiego internetu. NASK – Państwowy Instytut Badawczy, 2021.

[149] Reuters Institute for the Study of Journalism, Digital News Report 2019, Oxford: University of Oxford, 2019.

społeczeństwo rozpoznawania propagandy i *fake newsów*[150], co według niezależnych badań znacząco zmniejszyło podatność Finów na dezinformację. W Polsce edukacja medialna realizowana jest m.in. w ramach przedmiotu Wiedza o społeczeństwie i rozmaitych inicjatyw pozarządowych. Organizacje takie jak Fundacja Panoptykon, Stowarzyszenie Demagog czy projekt *FakeHunter* (Polska Agencja Prasowa we współpracy z *GovTech*) prowadzą szkolenia, kampanie informacyjne oraz dostarczają narzędzi do samodzielnej weryfikacji informacji. W 2019 r. powstał nawet poradnik dla administracji publicznej "*Jak zwalczać dezinformację*?"[151], zawierający praktyczne wskazówki reagowania i uodparniania społeczności na fałsz[152]. Media również włączają odbiorców do działań – np. platforma Demaskator24 uruchomiona przez Polska Press umożliwia zgłaszanie podejrzanych newsów, które są następnie sprawdzane przez dziennikarzy śledczych. Na poziomie międzynarodowym prowadzone są też kampanie prewencyjne (*pre-bunking*). Przykładowo, przed wyborami do Parlamentu Europejskiego w 2019 r. UE uruchomiła akcje informujące obywateli o typowych narracjach dezinformacyjnych pojawiających się w okresie wyborczym, aby zwiększyć ich czujność. Podobnie w trakcie pandemii COVID-19 Światowa Organizacja Zdrowia (WHO) i UNICEF organizowały kampanie uświadamiające o istnieniu infodemii – zalewu szkodliwych mitów – zachęcając do sprawdzania informacji zdrowotnych u ekspertów[153]. W 2023 r. platformy internetowe we współpracy z badaczami testowały wyświetlanie użytkownikom krótkich klipów edukacyjnych demaskujących techniki manipulacji (np. jak działa *clickbait*, jak rozpoznawać boty) – eksperymenty w Polsce i Czechach wykazały, że taka prewencyjna edukacja zwiększa odporność odbiorców na późniejszy

[150] Zob. Humprecht, E., Esser, F. i Van Aelst, P., Resilience to Online Disinformation: A Framework for Cross-National Comparative Research, The International Journal of Press/Politics, 25(3), 493-516.

[151] Zob. Ministerstwo Cyfryzacji, Jak zwalczać dezinformację? Poradnik dla administracji publicznej, Warszawa: MC, 2019.

[152] Zob. KRRiT, Fake news – dezinformacja online. Próby przeciwdziałania tym zjawiskom z perspektywy instytucji międzynarodowych oraz wybranych państw UE, w tym Polski, Warszawa, 2020.

[153] Zob. World Health Organization (WHO), Managing the COVID-19 Infodemic: Promoting Healthy Behaviours and Mitigating the Harm from Misinformation and Disinformation, Geneva: WHO, 2020.

kontakt z *fake newsem*[154]. Porównanie efektów i wyzwań. Dotychczasowe strategie różnią się podejściem, ale łączy je świadomość, że nie ma jednej cudownej recepty na dezinformację – potrzebne jest synergiczne działanie wielu środków. Kraje takie jak Francja i Niemcy postawiły na rozwiązania prawne, które szybko usztywniły ramy odpowiedzialności platform, lecz spotkały się z dyskusją o granice wolności słowa[155]. Inicjatywy samoregulacyjne (Włochy, UE) tworzą cenną przestrzeń współpracy branżowej, choć ich skuteczność zależy od dobrej woli uczestników i presji opinii publicznej.

Rozwój technologii wykrywających fałsz nabiera tempa, ale stoi przed nim wyzwanie – wyścig z równie dynamicznym rozwojem generatywnej AI. Edukacja społeczna przynosi efekty dopiero w dłuższej perspektywie, wymaga stałych nakładów i aktualizowania programów, by nadążać za nowymi trikami manipulatorów. W Polsce świadomość problemu na szczeblu władz i społeczeństwa wyraźnie wzrosła po 2020 r. (kryzysy dezinformacyjne wokół pandemii i wojny na Ukrainie). Podjęto szereg działań – od kampanii społecznych po ulepszanie prawa prasowego – jednak brak kompleksowej strategii anty-dezinformacyjnej z prawdziwego zdarzenia. Mimo to, polskie instytucje aktywnie uczestniczą w inicjatywach międzynarodowych[156] (np. *East StratCom Task Force* UE, współpraca z NATO w zakresie strategicznej komunikacji) i rozwijają krajowe zdolności (centrum analityczne NASK, piony ds. walki z dezinformacją w służbach).

Globalne doświadczenia wskazują, że najbardziej odporny na dezinformację jest ekosystem informacyjny, w którym prawo wspiera odpowiedzialność i transparentność, technologia pomaga wychwycić kłamstwo, media i platformy współpracują w dobrej wierze z weryfikatorami, a społeczeństwo posiada umiejętność krytycznego myślenia. Polska, podobnie jak inne państwa, stoi przed wyzwaniem zintegrowania tych elementów. Ze względu na położenie geopolityczne i bycie celem operacji informacyjnych (zwłaszcza ze wschodu), musi szczególnie intensywnie rozwijać mechanizmy obronne. Trzeba przy tym zachować równowagę – chronić przestrzeń

[154] Zob. J. Roozenbeek, J. i S. Van der Linden, Fake News Game Confers Psychological Resistance Against Online Misinformation, Palgrave Communications, nr 5, 2019.

[155] Bundesrepublik Deutschland, Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG), uchwalone 1 września 2017 r. (BGBl. I S. 3352), z późn. zm., ostatnia nowelizacja: art. 29 ustawy z dnia 6 maja 2024 r. (BGBl. 2024 I Nr. 149).

[156] Zob. NATO, NATO 2022 Strategic Concept, 2022.

informacyjną przed zalewem fałszu, nie dławiąc przy tym wolności słowa i pluralizmu debaty. Kontynuacja międzynarodowej współpracy (zarówno na forum UE, jak i NATO) w zakresie dzielenia się ostrzeżeniami o nowych kampaniach dezinformacyjnych, wspólnego rozwijania technologii i standardów postępowania – będzie kluczowa. Dezinformacja to zagrożenie o charakterze transgranicznym, dynamicznie ewoluujące wraz z mediami, dlatego odpowiedź na nie również musi być wielowymiarowa i elastyczna. Tylko takie kompleksowe podejście zwiększy odporność państwa i społeczeństwa na współczesne zagrożenia dezinformacyjne i medialne.

V

Technologie analizy treści audiowizualnych

5.1. Metody wykrywania zmanipulowanych treści audiowizualnych

W dobie rosnącej dezinformacji oraz dynamicznego rozwoju narzędzi cyfrowych, zdolność do wykrywania zmanipulowanych treści audiowizualnych staje się jednym z kluczowych wyzwań dla struktur odpowiedzialnych za bezpieczeństwo państwa i informacyjne. Zmanipulowane obrazy, filmy czy nagrania audio wykorzystywane są nie tylko do osiągania celów propagandowych, ale również do destabilizacji społecznej, podsycania konfliktów lub ingerencji w procesy polityczne. Konieczność przeciwdziałania tego typu zagrożeniom doprowadziła do wykształcenia się specjalistycznych metod i technologii wykorzystywanych w analizie treści audiowizualnych – w tym także rozwiązań wykorzystujących sztuczną inteligencję i uczenie maszynowe.

Cyfrowa analiza forensyczna (ang. *digital forensics*) stanowi jedno z głównych narzędzi w identyfikowaniu zmanipulowanych materiałów wizualnych i dźwiękowych. W tym kontekście stosuje się wiele technik umożliwiających wykrycie modyfikacji w strukturze pliku, zmian w metadanych lub nienaturalnych artefaktów wskazujących na potencjalne manipulacje. W szczególności analizie podlegają zjawiska takie jak duplikacja fragmentów obrazu – występująca, gdy fragmenty jednego zdjęcia są kopiowane i wklejane w inne miejsce (celem ukrycia niepożądanych elementów lub dodania nowych obiektów). Wykrycie tego typu manipulacji jest możliwe dzięki analizie podobieństwa wzorców pikselowych lub gradientów[157].

[157] Zob. I. Amerini, L. Ballan, R. Caldelli, A.Del Bimbo, G. Serra, A SIFT-based forensic method for copy–move attack detection and transformation recover, IEEE Transactions on Information Forensics and Security, 6(3), 2011, s. 1099-1110.

Innym sygnałem są anomalie w szumie i strukturze kompresji – cyfrowe obrazy i nagrania podlegają kompresji, co skutkuje powstawaniem charakterystycznych wzorców artefaktów; zaburzenia tych wzorców, np. niejednorodna dystrybucja szumu, mogą świadczyć o edycji materiału[158]. Analizie poddaje się również metadane EXIF – informacje o pliku (czas i lokalizacja wykonania, model aparatu itp.) osadzone w obrazach, czy filmach. Nieprawidłowości lub brak metadanych mogą stanowić przesłankę do dalszej weryfikacji autentyczności[159]. Techniki te, choć skuteczne, wymagają precyzyjnej kalibracji oraz często wsparcia ze strony specjalistów – zarówno w zakresie interpretacji wyników, jak i potwierdzania intencjonalnych manipulacji (fałszerstwa) w przypadku wykrycia

Uzupełnieniem klasycznej analizy forensycznej jest metoda wstecznego wyszukiwania obrazem (ang. *reverse image search*), wykorzystywana głównie w walce z tzw. prostymi fałszywkami (ang. *cheap fakes*). Polega ona na porównaniu badanego obrazu lub klatki filmu z dostępnymi w Internecie zasobami w celu zidentyfikowania jego oryginalnego kontekstu. Jeżeli rzekomo nowe zdjęcie pojawiło się już wcześniej w innym miejscu lub czasie, można wykazać celowe wprowadzenie odbiorców w błąd co do jego znaczenia. Tę technikę stosuje się np. przy demaskowaniu manipulacji polegających na prezentowaniu starego zdjęcia z innego kraju jako aktualnej relacji z miejsca konfliktu. Narzędzia takie jak *Google Images* czy specjalistyczne wtyczki (np. InVID dla dziennikarzy) potrafią automatycznie podzielić nagranie wideo na klatki i przeszukać Internet w celu znalezienia ich źródła. Projekt InVID (finansowany przez UE) dostarcza cały zestaw tego typu narzędzi – określany wręcz jako „szwajcarski scyzoryk weryfikacji wideo" – dostępnych jako rozszerzenie przeglądarki. Oferuje on m.in. analizę kontekstu publikacji (źródło, data, powiązane komentarze), dekompozycję filmu na klatki kluczowe z integracją wyszukiwarek obrazów, ekstrakcję metadanych, a nawet filtry forensyczne do wstępnego wykrywania ingerencji w obraz. Takie algorytmiczne narzędzia stały się standardowym wyposażeniem zespołów *fact-checkingowych* na świecie i okazały się skuteczne w demaskowaniu większości zmanipulowanych treści krążących w sieci.

[158] Zob. J. Fridrich, D. Soukal, J. Lukáš, Detection of copy–move forgery in digital images, Proceedings of the Digital Forensic Research Workshop, 2003.

[159] Zob. A. Jain, Reverse Image Search: How and Why It Works, Journal of Visual Communication and Image Representation, 60, 170-179.

W przypadku nagrań audio, analogicznie do obrazów, możliwe jest stosowanie technik forensycznych takich jak analiza tła dźwiękowego, widma częstotliwości czy detekcja nienaturalnych przerw w sygnale. Zmanipulowane nagrania – w szczególności tzw. *deepfake audio* – często zawierają subtelne nieciągłości akustyczne, które mogą wskazywać na ich syntetyczne pochodzenie[160]. Do analizy wykorzystuje się m.in. spektrogramy (wizualizacje składowych częstotliwości dźwięku), detekcję niespójności rytmu oddechu lub odgłosów tła, a także identyfikację artefaktów pozostawianych przez syntezatory mowy. Zastosowanie tych metod pozwala wykryć manipulacje, które w przeciwnym razie mogłyby pozostać niezauważone – szczególnie w kontekście politycznym lub militarnym, gdzie autentyczność komunikatów odgrywa kluczową rolę.

Coraz częściej narzędzia analityczne do wykrywania fałszerstw audiowizualnych wspierane są przez algorytmy sztucznej inteligencji, w szczególności przez głębokie sieci neuronowe. Systemy uczące się na ogromnych zbiorach danych potrafią rozpoznawać wzorce typowe dla materiałów zmanipulowanych. Przykładowo, sieci konwolucyjne wykorzystywane są do identyfikacji nienaturalnych deformacji w strukturze obrazu, błędów cieniowania czy niespójności oświetlenia[161]. Z kolei sieci rekurencyjne (RNN) lub modele transformatorowe mogą analizować ciągi klatek wideo, wychwytując nienormalne zależności czasowe – na przykład brak naturalnego synchronu ruchu ust z dźwiękiem czy niefizjologiczne częstotliwości mrugnięć powiek w wygenerowanych nagraniach. Literatura opisuje liczne szczególne sygnały, które mogą zdradzać *deepfake* – od nienaturalnego mrugania (gdy postać na wideo mruga rzadziej lub inaczej niż typowy człowiek) po nieregularności w odbiciu światła w oczach lub delikatne artefakty na krawędziach obiektów po kompozycji obrazu[162]. W praktyce najwyższą skuteczność osiąga się przez podejście hybrydowe, łączące tradycyjną wiedzę ekspercką (dotyczącą typowych oznak montażu) z mocą nowoczesnych

[160] Zob. Y. Zhang, X. Ma, X. Wang, J. Sun, Detecting Audio Deepfakes Using Frame-Level Self-Attention, IEEE Transactions on Dependable and Secure Computing, 18(5), 2415-2428.

[161] Zob. T.T. Nguyen, C.M. Nguyen, D.T. Nguyen, D.T., Nguyen, S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573, 2019.

[162] Zob. Y. Li, M.-C. Chang, S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong 2018, s. 1-7.

sieci neuronowych – takie połączenie zwiększa wykrywalność nawet złożonych manipulacji[163]. Coraz większy nacisk kładzie się również na analizę multimodalną, łączącą cechy wizualne, dźwiękowe, a często i tekstowe. System bada jednocześnie warstwę wideo (obrazy klatka po klatce) oraz warstwę audio (głos, tło) – a niekiedy również automatycznie wyodrębniony tekst mówiony – w poszukiwaniu niespójności między tymi elementami. Przykładowo, obraz postaci może nie zgadzać się z przypisanym jej głosem lub treść wypowiedzi kolidować z widocznym kontekstem. Zaawansowane algorytmy potrafią wykrywać mikroekspresje twarzy niepasujące do tonu głosu czy inne drobne niespójności między modalnościami, by skuteczniej ujawnić fałszerstwo. Bbadania potwierdzają, że fuzja cech z obrazu i dźwięku znacząco zwiększa wykrywalność fałszywych nagrań względem metod opartych tylko na jednej modalności[164]. W ramach programów badawczych, takich jak inicjatywy DARPA, opracowano nawet kompleksowe algorytmy wielomodalne zdolne do jednoczesnej analizy tekstu, audio i wideo w celu identyfikacji syntetycznych mediów. Przykładowo firma Kitware, uczestnicząca w projekcie *SemaFor* (*Semantic Forensics*), stworzyła zestaw narzędzi potrafiących wykrywać różne typy manipulacji obrazu i dźwięku oraz przypisywać je do konkretnych technik generatywnych[165].

Pomimo dynamicznego postępu technologii, wykrywanie zmanipulowanych treści audiowizualnych pozostaje zadaniem trudnym, wymagającym integracji wielu metod oraz ciągłej aktualizacji narzędzi. Kluczowymi problemami są m.in.: rosnąca jakość generowanych materiałów syntetycznych (np. obrazy i wideo tworzone za pomocą sieci GAN są coraz trudniejsze do odróżnienia od autentycznych materiałów) oraz brak standardów walidacji i certyfikacji treści cyfrowych. Istotne pozostają także wyzwania związane z interpretacją kontekstu – kulturowego, społecznego, politycznego – danego materiału: sama detekcja technicznej manipulacji nie zawsze pozwala stwierdzić, czy treść ma intencjonalnie wprowadzać w błąd. Z tego względu coraz większą rolę odgrywa nie tylko technologia, ale także edukacja medialna, współpraca międzynarodowa oraz tworzenie narzędzi

[163] Zob. M. Westerlund, The Emergence of Deepfake Technology: A Review, Technology Innovation Management Review, 9(11), 39-52.

[164] Zob. T.T. Nguyen, C.M. Nguyen, D.T. Nguyen, D.T., Nguyen, S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573, 2019.

[165] https://www.kitware.com/darpa-enlists-kitware-on-program-to-detect-falsified-media/ [Dostęp: 28 marca 2024].

umożliwiających obywatelom weryfikację treści. Współczesne państwo, które pragnie skutecznie chronić swoją przestrzeń informacyjną, powinno inwestować zarówno w nowoczesne technologie detekcyjne, jak i w rozwój kompetencji analitycznych swoich służb oraz społeczeństwa.

5.2. Algorytmy analizy obrazu i dźwięku w detekcji dezinformacji

Wykrywanie dezinformacji rozpowszechnianej za pomocą materiałów wizualnych i dźwiękowych wymaga zastosowania wyspecjalizowanych algorytmów analizy obu tych typów danych. W praktyce oznacza to połączenie metod komputerowego rozpoznawania obrazu (ang. *computer vision*) oraz przetwarzania sygnałów audio, wspartych technikami uczenia maszynowego. Poniżej omówiono wybrane klasy algorytmów wykorzystywanych do detekcji fałszerstw w mediach wizualnych i dźwiękowych, ze szczególnym uwzględnieniem ich roli w identyfikowaniu kampanii dezinformacyjnych. Algorytmy analizy obrazu koncentrują się na wychwytywaniu cech świadczących o manipulacji zdjęciem lub klatką wideo. Tradycyjne podejścia opierały się na metodach inżynierii cech – eksperci wyodrębniali charakterystyczne właściwości obrazu (np. histogramy barw, statystyki szumu, wektory ruchu) i trenowali klasyfikatory w celu odróżniania obrazów prawdziwych od fałszywych. Obecnie jednak dominują algorytmy oparte na głębokich sieciach neuronowych konwolucyjnych CNN, zdolnych do automatycznej ekstrakcji istotnych cech z danych. Wyspecjalizowane architektury CNN (takie jak *XceptionNet* czy *EfficientNet*) zostały z powodzeniem zastosowane do detekcji *deepfake* – sieci te uczą się subtelnych anomalii na poziomie pikseli pozostawianych przez proces generowania fałszywego obrazu np. nieregularnych artefaktów kompresji czy nienaturalnego rozmycia na krawędziach wklejonej twarzy). Wyniki konkursu *Facebook Deepfake Detection Challenge*, w którym najskuteczniejsze modele osiągnęły skuteczność klasyfikacji na poziomie około 65%, unaoczniają skalę trudności związanej z wykrywaniem treści syntetycznych[166]. Wskaźnik ten, bliski losowemu przyporządkowaniu, wskazuje na istotne ograniczenia aktualnych

[166] U.S. Government Accountability Office (GAO), Science & Tech Spotlight: Combating Deepfakes, GAO-24-107292, Waszyngton, 2024.

metod detekcji *deepfake'ów* oraz potwierdza, że obszar ten nadal stanowi wyzwanie badawcze i technologiczne[167]. Analizy przeprowadzone na zbiorach błędnie sklasyfikowanych materiałów wykazały, że modele nie opierały się wyłącznie na identyfikacji jednoznacznych artefaktów pochodzenia syntetycznego – takich jak charakterystyczne zniekształcenia pikselowe czy regularności geometryczne – lecz raczej na wykrywaniu rozproszonych i trudnych do formalizacji anomalii w strukturze obrazu. Sugeruje to, że sieci neuronowe funkcjonowały w sposób zbliżony do percepcji ludzkiej, identyfikując ogólną niespójność między obrazem a oczekiwanym wzorcem rzeczywistości, przy czym czyniły to z większą precyzją w odniesieniu do mikrostruktur obrazu.

W kontekście analizy wideo, coraz większe znaczenie zyskują algorytmy uwzględniające zależności czasowe między kolejnymi klatkami nagrania. W szczególności, istotne są metody analizujące płynność i koherencję sekwencji ruchów – takich jak mruganie powiek, ruchy ust, czy mimika twarzy – które w syntetycznych materiałach często wykazują nienaturalne zachowania, odbiegające od wzorców biologicznych charakterystycznych dla rzeczywistych nagrań. Wczesne generacje *deepfake'ów* często charakteryzowały się nieregularnościami w synchronizacji artykulacji z dźwiękiem, co wykorzystywano przy konstrukcji detektorów opartych na analizie przepływu optycznego (ang. *optical flow*). Algorytmy te identyfikowały m.in. skokowe przemieszczenia fragmentów obrazu, zaburzenia wektorów ruchu lub asynchronię między modalnościami wizualną i akustyczną.

Obecnie stosowane podejścia wykorzystują modele hybrydowe, integrujące konwolucyjne sieci neuronowe, służące do analizy statycznych właściwości klatek wideo, z rekurencyjnymi sieciami neuronowymi lub architekturami transformatorowymi, odpowiedzialnymi za modelowanie ciągłości czasowej. Tego typu konstrukcje pozwalają na wykrywanie subtelnych niespójności temporalnych, takich jak mikroróżnice w synchronizacji wizualnej reprezentacji artykulacji z emitowanym dźwiękiem, czy też nieadekwatność ekspresji emocjonalnej względem intonacji głosu. Analiza multimodalna oparta na takich modelach umożliwia skuteczniejsze rozpoznawanie treści syntetycznych, których identyfikacja byłaby trudna lub niemożliwa przy zastosowaniu pojedynczej modalności. Badania potwierdzają,

[167] Zob. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang and C.C. Ferrer, The DeepFake Detection Challenge (DFDC) Dataset, arXiv preprint arXiv:2006.07397, 2020.

że fuzja cech z obu modalności znacząco zwiększa skuteczność detekcji fałszywych nagrań w porównaniu do klasycznych metod opartych wyłącznie na obrazie lub tylko na dźwięku[168]. Algorytmy analizy sygnału dźwiękowego odgrywają istotną rolę w procesie detekcji treści zmanipulowanych, zwłaszcza w kontekście stale udoskonalanych metod syntezy mowy, które pozwalają na generowanie przekonujących, lecz fałszywych nagrań głosowych. W analogii do rozwiązań stosowanych w analizie obrazu, również w przetwarzaniu sygnału akustycznego coraz częściej wykorzystuje się architektury głębokiego uczenia, przede wszystkim konwolucyjne sieci neuronowe oraz sieci rekurencyjne, takie jak LSTM (ang. *Long Short-Term Memory*). Modele te operują najczęściej na spektrogramach – graficznych reprezentacjach widma częstotliwości sygnału w funkcji czasu – co umożliwia uchwycenie istotnych cech fonetycznych i temporalnych głosu. Celem tych algorytmów jest wykrywanie subtelnych aberracji w strukturze sygnału mowy, które nie są bezpośrednio słyszalne dla człowieka, lecz stanowią wskaźnik syntetycznego pochodzenia nagrania. Przykładowo, generowane przez algorytmy *text-to-speech* głosy mogą cechować się brakiem naturalnych niuansów intonacyjnych, jednolitym widmem bez typowego dla nagrań realnych szumu tła czy mikrodrgań w częstotliwości. Współczesne badania z zakresu wykrywania tzw. *audio-deepfake'ów* skupiają się na identyfikacji charakterystycznych właściwości sygnału – takich jak niestandardowe współczynniki *cepstralne* (MFCC), anomalne rozkłady energii w wyższych pasmach częstotliwości czy nieciągłości w fazie sygnału – które są trudne do odwzorowania przez systemy syntezy mowy[169].

W celu zwiększenia odporności systemów detekcji na różnorodność źródeł oraz technik generatywnych, rozwijane są także metody oceny odporności modeli na zmienność językową i fonetyczną. W tym kontekście organizowane są wyspecjalizowane konkursy i benchmarki – takie jak *ASVspoof* czy *ADD Challenge* – których celem jest stymulowanie badań nad algorytmami detekcji odpornymi na manipulacje głosowe niezależnie od języka, akcentu czy jakości nagrania.

[168] Zob. T.T. Nguyen, C.M. Nguyen, D.T. Nguyen, D.T., Nguyen, S. Nahavandi, Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573, 2019.

[169] U.S. Government Accountability Office (GAO), Science & Tech Spotlight: Combating Deepfakes, GAO-24-107292, Waszyngton, 2024.

Fałszowanie sygnału dźwiękowego może przybierać różne formy, począwszy od całkowicie syntetyzowanej mowy generowanej na podstawie tekstu (ang. *text-to-speech*), poprzez techniki przekształcania głosu jednej osoby w głos innej (ang. *voice conversion*), aż po częściowe manipulacje, takie jak podmiana pojedynczych słów w autentycznym nagraniu. Skuteczne algorytmy detekcyjne muszą być zatem zdolne nie tylko do identyfikacji pełnych *deepfake'ów*, ale również do wychwytywania fragmentarycznych fałszerstw osadzonych w prawdziwym materiale. W praktyce stosuje się w tym celu klasyfikatory uczenia maszynowego trenowane na dużych zbiorach nagrań – zarówno autentycznych, jak i sztucznych. Modele oparte na architekturach takich jak *ResNet* czy *EfficientNet* osiągają w kontrolowanych warunkach laboratoryjnych wysokie wskaźniki skuteczności, nierzadko przekraczające 90% trafności klasyfikacji na znanych zbiorach danych. Niemniej, analogicznie jak w przypadku detekcji obrazów, problemem pozostaje ograniczona zdolność generalizacji. Wystarczy, aby system napotkał nagranie wygenerowane nową, dotąd nieznaną metodą lub zawierające głos osoby, której charakterystyka fonetyczna nie występowała w danych treningowych, a skuteczność klasyfikatora może ulec istotnemu obniżeniu. Z tego względu współczesne badania eksplorują również metody uczenia transferowego oraz uczenia niezależnego od domeny (ang. *domain-invariant learning*), które mają na celu nauczenie modeli rozpoznawania ogólnych wzorców sztuczności mowy, a nie tylko dopasowania do konkretnych przykładów. Przykładowo, zamiast rozpoznawać charakterystyczne cechy syntetycznych głosów używanych w znanych systemach, modele uczone na poziomie meta-cech dążą do uchwycenia uniwersalnych właściwości odróżniających mowę naturalną od generowanej – takich jak nieregularna dynamika sygnału, przypadkowe przesunięcia czasowe lub fizjologiczna nieciągłość oddechowa.

Podsumowując, algorytmy analizy dźwięku – obok wizualnych – stanowią integralny komponent systemów wykrywania dezinformacji audiowizualnej. Od metod forensycznych, opartych na analizie struktury sygnału, przez głębokie modele uczące się z danych, aż po złożone systemy multimodalne integrujące wiele źródeł informacji – rozwiązania te nieustannie ewoluują w odpowiedzi na zmieniające się zagrożenia. Ich skuteczne wdrożenie w praktyce operacyjnej – zarówno przez instytucje państwowe, jak i podmioty prywatne – ma fundamentalne znaczenie dla wczesnego wykrywania

manipulacji oraz ograniczania ich potencjalnego wpływu na odbiorców. Współcześnie technologie te coraz częściej wchodzą w skład zintegrowanych platform bezpieczeństwa informacyjnego, w których identyfikacja treści zmanipulowanych odbywa się automatycznie i w skali masowej.

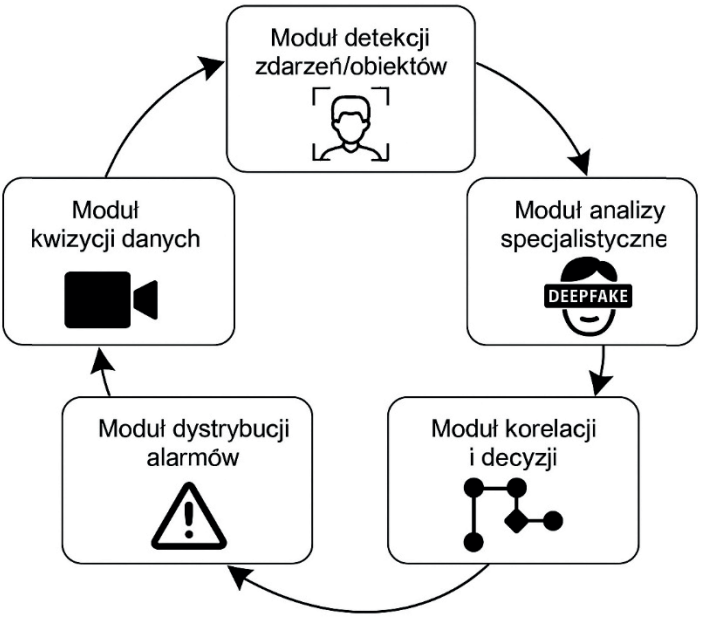
5.3. Systemy monitorowania wideo w czasie rzeczywistym

Współczesne zagrożenia związane z dezinformacją, manipulacjami informacyjnymi oraz coraz bardziej wyrafinowanymi technologiami syntetyzującymi treści multimedialne wymagają zastosowania nowoczesnych i wysoce responsywnych narzędzi analitycznych. W tym kontekście szczególnego znaczenia nabierają systemy monitorowania treści wideo w czasie rzeczywistym, których zadaniem jest nie tylko rejestracja i analiza materiału audiowizualnego, ale również dynamiczne wykrywanie zagrożeń oraz natychmiastowe generowanie alertów ostrzegawczych. Celem tych systemów jest wspieranie bezpieczeństwa publicznego, infrastrukturalnego i informacyjnego poprzez umożliwienie szybkiego reagowania na incydenty – zarówno fizyczne, jak i informacyjne – w warunkach silnie nasyconych danymi.

Systemy tego typu funkcjonują w zróżnicowanych środowiskach operacyjnych, zależnie od charakterystyki analizowanej przestrzeni i rodzaju potencjalnych zagrożeń. Po pierwsze, znajdują zastosowanie w monitoringu bezpieczeństwa fizycznego – np. w systemach nadzoru miejskiego (CCTV), ochrony infrastruktury krytycznej, obiektów strategicznych czy masowych imprez publicznych. W tych przypadkach systemy analizują sygnały wideo pochodzące z sieci kamer w celu identyfikacji zagrożeń takich jak: wtargnięcia na obszary zastrzeżone, pozostawienie bagażu bez nadzoru, bójki, akty wandalizmu czy inne formy podejrzanego zachowania. Dodatkowo coraz częściej integruje się te systemy z czujnikami audio, co umożliwia detekcję wystrzałów, szkła tłuczonego czy nagłych okrzyków, co jeszcze bardziej zwiększa ich funkcjonalność. Po drugie, systemy te pełnią istotną funkcję w sferze informacyjnej, gdzie wykorzystywane są do bieżącego monitorowania treści publikowanych na platformach społecznościowych, w serwisach *streamingowych* oraz w tradycyjnych kanałach medialnych. Ich zadaniem jest wykrywanie manipulacji medialnych, w tym fałszywych

nagrań wideo (np. *deepfake*), fragmentów propagandy wizualnej, materiałów o nieznanym pochodzeniu lub sygnaturach wskazujących na ingerencję cyfrową. Część z tych rozwiązań działa w trybie stałym *(*ang. *non-stop*), przeszukując ogromne ilości danych w poszukiwaniu tzw. treści anormalnych – czyli odchyleń od znanych wzorców poznanych w procesie trenowania algorytmów.

Z technologicznego punktu widzenia, systemy monitorujące w czasie rzeczywistym opierają się na zaawansowanej architekturze przetwarzania strumieni danych. Strumień wideo (np. z kamer miejskich, dronów, monitoringu przemysłowego czy transmisji *live-streaming*) jest segmentowany na klatki lub pakiety danych i analizowany w czasie rzeczywistym przez kaskadowe moduły algorytmiczne. Typowy potok przetwarzania przedstawiono na rysunku 2.



Rysunek 2. Schemat przetwarzania treści audiowizualnych w czasie rzeczywistym
Źródło: opracowanie własne.

System może obejmować następujące komponenty:

1. Moduł akwizycji danych – odpowiada za odbiór strumienia wideo w czasie rzeczywistym oraz jego wstępną obróbkę, w tym redukcję szumów, kompresję lub standaryzację rozdzielczości;

2. Moduł detekcji zdarzeń i obiektów – realizuje funkcje wykrywania istotnych elementów w obrazie, takich jak osoby, pojazdy, zgromadzenia, gesty, z wykorzystaniem algorytmów opartych na sieciach neuronowych typu YOLO, SSD czy *EfficientDet*;
3. Moduł analizy specjalistycznej – wdraża bardziej zaawansowane procedury analizy, np. wykrywanie treści syntetycznych (*deepfake*), rozpoznawanie twarzy, detekcja broni, rozpoznawanie symboliki ekstremistycznej lub śledzenie trajektorii ruchu;
4. Moduł korelacji i decyzji – integruje dane z kolejnych klatek wideo oraz z innych źródeł (np. czujników akustycznych, informacji kontekstowych), umożliwiając wnioskowanie i klasyfikację sytuacji jako potencjalnie niebezpiecznej lub wymagającej interwencji;
5. Moduł dystrybucji alarmów – odpowiedzialny za emisję powiadomień do operatorów centrum monitoringu, jednostek reagowania kryzysowego lub systemów automatycznego przeciwdziałania, np. zablokowania transmisji.

Całość systemu musi charakteryzować się minimalnym opóźnieniem przetwarzania – rzędu ułamków sekundy – co ma zasadnicze znaczenie dla skuteczności operacyjnej, zwłaszcza w kontekście materiałów wideo transmitowanych na żywo. Tylko tak krótki czas reakcji umożliwia wstrzymanie szkodliwego przekazu, zanim dotrze on do szerokiego grona odbiorców i wywoła rzeczywiste skutki społeczne lub polityczne.

Z punktu widzenia architektury systemowej, szczególną rolę odgrywa obecnie tzw. *edge computing*, czyli rozproszone przetwarzanie danych blisko źródła ich generacji. Przykładowo, inteligentne kamery przemysłowe mogą być wyposażone w lokalne moduły AI, które wstępnie analizują strumień wideo i przesyłają do centrum jedynie dane uznane za istotne (np. detekcję obecności broni lub twarzy z listy osób poszukiwanych). Pozwala to ograniczyć przesył danych, przyspieszyć reakcję oraz odciążyć infrastrukturę sieciową.

W kontekście przeciwdziałania dezinformacji coraz częściej implementuje się prototypowe systemy wczesnego ostrzegania, jak MEWS (ang. *Misinformation Early Warning System*), które poza analizą samych treści monitorują również dynamikę ich rozpowszechniania. System taki może wykryć, że zmanipulowany obraz lub film jest intensywnie udostępniany w określonej grupie społecznościowej, co może świadczyć o zorganizowanej

kampanii wpływu. Dzięki analizie grafów interakcji użytkowników i źródeł publikacji możliwe jest wczesne zidentyfikowanie koordynowanych działań dezinformacyjnych oraz ich powiązań z konkretnymi podmiotami. Systemy tego typu są również wdrażane przez platformy cyfrowe, które w czasie rzeczywistym analizują przesyłane materiały – zarówno wizualne, jak i dźwiękowe – pod kątem obecności znanych fałszywych treści. Przykładowo, Facebook i YouTube stosują detektory *deepfake'ów* oraz filtry identyfikujące przemoc fizyczną czy pornografię. Transmisja materiału może zostać wstrzymana do czasu zakończenia weryfikacji przez moderatora. Co istotne, decyzje o blokadzie materiału nie są podejmowane wyłącznie przez algorytmy – w większości przypadków ostateczne rozstrzygnięcie należy do zespołu przeszkolonych analityków.

W obliczu narastających zagrożeń związanych z dezinformacją oraz operacjami wpływu prowadzonymi przez podmioty państwowe i niepaństwowe, instytucje państwowe oraz organizacje międzynarodowe, takie jak NATO, Unia Europejska (UE) czy Organizacja Narodów Zjednoczonych (ONZ), intensyfikują działania mające na celu monitorowanie treści audiowizualnych w czasie rzeczywistym. Celem tych inicjatyw jest identyfikacja, analiza oraz neutralizacja fałszywych informacji, które mogą destabilizować społeczeństwa, podważać zaufanie do instytucji demokratycznych oraz wpływać na procesy polityczne i społeczne. Jednym z kluczowych podmiotów zaangażowanych w przeciwdziałanie dezinformacji jest Centrum Eksperckie NATO ds. Komunikacji Strategicznej (NATO *Strategic Communications Centre of Excellence, StratCom COE)* z siedzibą w Rydze, Łotwa[170]. *StratCom COE* koncentruje się na analizie metodologii stosowanych przez podmioty prowadzące operacje dezinformacyjne, ze szczególnym naciskiem na środowisko cyfrowe oraz nowe technologie. Centrum prowadzi badania nad wykorzystaniem mediów społecznościowych jako narzędzia w wojnie hybrydowej oraz analizuje implementację polityki komunikacji strategicznej NATO. W przypadku identyfikacji zmanipulowanych treści, *StratCom* COE jest w stanie szybko uruchomić procedury komunikacji kryzysowej, sprostowań lub działań dyplomatycznych, co ma kluczowe znaczenie w minimalizowaniu skutków dezinformacji. Unia Europejska również podejmuje zdecydowane kroki w walce z dezinformacją. Platforma *EUvsDisinfo*,

[170] https://www.act.nato.int/article/stratcom-coe-2024/ [Dostęp: 28 marca 2024].

uruchomiona po nielegalnej aneksji Krymu przez Rosję w 2014 roku, zajmuje się wykrywaniem, analizą oraz podnoszeniem świadomości na temat kampanii dezinformacyjnych wymierzonych w UE[171]. Platforma ta gromadzi i publikuje przykłady fałszywych narracji, dostarczając obywatelom narzędzi do rozpoznawania i przeciwdziałania dezinformacji. W kontekście krajowym, Polska również podejmuje działania mające na celu przeciwdziałanie dezinformacji. Zespół ds. Dezinformacji przy Komisji badającej wpływy rosyjskie i białoruskie opublikował raport analizujący działania dezinformacyjne ze strony Rosji i Białorusi oraz oceniający skuteczność podjętych środków zaradczych. Raport zawiera również rekomendacje dotyczące strategii i działań mających na celu zwiększenie odporności państwa na tego rodzaju zagrożenia[172].

Istotnym aspektem strategii przeciwdziałania zagrożeniom informacyjnym, zwłaszcza dezinformacji, jest współpraca międzynarodowa. Charakter współczesnych kampanii dezinformacyjnych, prowadzonych często przez podmioty państwowe lub ponadnarodowe, wymaga bowiem skoordynowanego działania instytucji międzynarodowych, takich jak NATO, Unia Europejska (UE) czy Organizacja Narodów Zjednoczonych (ONZ). W ramach tych struktur realizowane są liczne inicjatywy obejmujące wymianę informacji wywiadowczych, prowadzenie wspólnych ćwiczeń, rozwijanie procedur operacyjnych oraz wypracowywanie standardów prawnych i etycznych regulujących sposoby przeciwdziałania manipulacjom treści audiowizualnych i informacyjnych. Przykładem takiej współpracy jest koordynacja działań pomiędzy NATO Strategic *Communications Centre of Excellence* (*StratCom COE*), a *East StratCom Task Force* – specjalną komórką utworzoną przez Unię Europejską w 2015 roku, której głównym zadaniem jest wykrywanie oraz zwalczanie rosyjskich kampanii dezinformacyjnych wymierzonych w kraje europejskie. W ramach tej współpracy obie jednostki regularnie wymieniają się analizami dotyczącymi nowych technik dezinformacyjnych, prowadzą wspólne warsztaty oraz szkolenia dla analityków i specjalistów ds. komunikacji strategicznej. Wspólne inicjatywy obejmują także symulacje operacyjne, których celem jest doskonalenie procedur szybkiego reagowania kryzysowego w sytuacji pojawienia się zmanipulowanych przekazów

[171] https://euvsdisinfo.eu/ [Dostęp: 24 marca 2024].

[172] https://www.gov.pl/web/sprawiedliwosc/raport-zespolu-ds-dezinformacji-komisji-ds-badania-wplywow-rosyjskich-i-bialoruskich [Dostęp: 28 marca 2024].

mogących destabilizować opinię publiczną lub wpływać na procesy decyzyjne[173]. Ponadto Unia Europejska wraz z NATO angażują się w rozwój procedur oraz technicznych standardów identyfikowania, raportowania oraz przeciwdziałania dezinformacji. W ramach takiego współdziałania powstały dokumenty strategiczne określające zakres odpowiedzialności oraz mechanizmy współpracy poszczególnych państw członkowskich w przypadku wystąpienia masowych kampanii dezinformacyjnych. Przykładowo, *„EU--NATO Joint Declaration on Cooperation*" określa wytyczne współpracy na poziomie politycznym, operacyjnym i technicznym w celu przeciwdziałania zagrożeniom hybrydowym, w tym także kampaniom dezinformacyjnym[174]. W ramach ONZ również podejmowane są działania mające na celu koordynację globalnych wysiłków w zakresie zwalczania dezinformacji, szczególnie w kontekście globalnych kryzysów, takich jak pandemia COVID-19 czy konflikty zbrojne. Jednym z przykładów jest inicjatywa *Verified* – kampania informacyjna prowadzona przez ONZ, która koncentruje się na walce z dezinformacją poprzez dostarczanie sprawdzonych informacji oraz współpracę z platformami społecznościowymi w zakresie oznaczania treści wprowadzających w błąd[175]. W praktyce współpraca ta obejmuje również techniczne aspekty wymiany informacji, w tym korzystanie ze wspólnych platform analitycznych umożliwiających błyskawiczne raportowanie oraz śledzenie rozprzestrzeniania się fałszywych treści. Organizowane są także cykliczne konferencje oraz seminaria z udziałem ekspertów ds. cyberbezpieczeństwa, komunikacji strategicznej, analityków danych oraz specjalistów ds. zarządzania kryzysowego, których celem jest wypracowanie wspólnych rekomendacji oraz najlepszych praktyk w przeciwdziałaniu dezinformacji na poziomie międzynarodowym[176]. Współpraca międzynarodowa w zakresie przeciwdziałania zagrożeniom informacyjnym ma zatem kluczowe znaczenie ze względu na globalny charakter współczesnych operacji wpływu. Efektywność tych działań jest uzależniona od sprawnej wymiany informacji,

[173] European External Action Service, Questions and Answers about the East StratCom Task Forc, European Union, 2023.

[174] NATO, EU-NATO cooperation – Factsheet, 2023.

[175] United Nations, Verified Campaign, UN Department of Global Communications, 2020.

[176] Zob. S. Drosio, S. Stanek, Big Data jako źródło informacji rozszerzające funkcjonowanie systemów wspomagania decyzji w zarządzaniu kryzysowym, Zeszyty Naukowe Politechniki Częstochowskiej, 26, 2017.

zharmonizowanych procedur oraz koordynacji operacyjnej pomiędzy instytucjami o zasięgu międzynarodowym.

Podsumowując, systemy monitorowania treści audiowizualnych w czasie rzeczywistym odgrywają kluczową rolę w strategiach obronnych instytucji państwowych oraz organizacji międzynarodowych. Dzięki nim możliwe jest szybkie identyfikowanie i neutralizowanie zagrożeń związanych z dezinformacją, co przyczynia się do ochrony integralności procesów demokratycznych oraz bezpieczeństwa informacyjnego społeczeństw. Systemy monitorowania wideo w czasie rzeczywistym zaś, to złożone i wysoce wyspecjalizowane układy, których znaczenie dla bezpieczeństwa narodowego rośnie wraz z eskalacją zagrożeń cyfrowych. Łącząc wydajność algorytmiczną z architekturą big data oraz mechanizmami decyzyjnymi, stanowią one nie tylko narzędzie reagowania na incydenty, ale również element strategicznego zarządzania informacją w erze konfrontacji hybrydowej i dezinformacyjnej.

5.4. Detekcja *deepfake* – skuteczność i ograniczenia

Detekcja *deepfake* stała się kluczowym obszarem badań ze względu na rosnące zagrożenie, jakie generowane syntetycznie obrazy, wideo i audio stanowią dla sfery informacyjnej. Współczesne metody wykrywania *deepfake'ów* opierają się głównie na zaawansowanych algorytmach uczenia maszynowego analizujących subtelne artefakty i niespójności pozostawiane przez algorytmy generujące fałszywe media[177]. W przypadku wideo najczęściej stosuje się konwolucyjne sieci neuronowe (CNN) do analizy kolejnych klatek filmu i wychwytywania anomalii w obrazie twarzy. Przykładowo, badacze potrafią wykrywać nienaturalne deformacje na krawędziach nakładanej twarzy, zaburzenia tekstury skóry czy artefakty pikselowe wokół oczu i ust, które mogą świadczyć o podmianie twarzy. Wiele metod koncentruje się na newralgicznych obszarach twarzy – m.in. ruchu ust, mruganiu powiek

[177] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, T.S. Bonaventura, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, S. Mandelli, G.L. Marcialis, M. Micheletto, A. Montibeller, G. Orrù, A. Ortis, P. Perazzo, D. Salvi, S. Tubaro, C.M. Tonti, M. Villari, D. Vitulano, Deepfake Media Forensics: State of the Art and Challenges Ahead, arXiv preprint arXiv:2408.00388v1, 2024.

czy refleksach światła w oczach – ponieważ tam często ujawniają się niedoskonałości algorytmów generujących *deepfake*. Na przykład we wczesnych generacjach *deepfake* zauważano brak regularnego mrugania u postaci, co stanowiło czytelny sygnał fałszerstwa. Obecnie jednak twórcy fałszywek wyeliminowali ten mankament i nowoczesne fałszywe nagrania zawierają już naturalnie wyglądające mrugnięcia[178]. Bardziej wyszukane techniki detekcji sięgają po cechy biologiczne – np. analizę pulsu na podstawie drobnych zmian koloru skóry twarzy. Badania finansowane przez DARPA w ramach programu MediFor opracowały prototypy wykrywające mikrodrgania naczyń krwionośnych na czole osoby z nagrania, co ujawnia obecność żywego pulsu lub jego brak w przypadku generowanej postaci. Równolegle rozwijane są metody wykorzystujące spójność międzyklatkową – sieci neuronowe LSTM lub modele uwzględniające ruch w wideo potrafią wyłapać nienaturalne przejścia między klatkami (np. „skoki" w pozycji głowy, które nie występują u autentycznej osoby). W przypadku audio (sztucznie generowanej mowy) stosuje się analogiczne podejścia oparte na uczeniu głębokim, dostosowane do sygnału dźwiękowego. Algorytmy analizują widmo częstotliwości głosu (np. spektrogramy) w poszukiwaniu anomalii – mogą to być nienaturalnie płaskie intonacje, artefakty cyfrowe w wysokich częstotliwościach lub brak charakterystycznych dla ludzkiego głosu cech, takich jak mikrodrgania częstotliwości podstawowej. Zastosowanie znajdują modele klasyfikujące audio (jak CNN lub modele rekurencyjne) uczone na obszernych zbiorach nagrań prawdziwych i fałszywych głosów. Mimo coraz doskonalszych syntetycznych głosów, istnieją detektory wychwytujące drobne nieciągłości mowy lub niewidoczne dla ludzkiego ucha różnice w charakterystyce akustycznej. Warto podkreślić, że ludzka percepcja bywa zawodna – w jednym z badań uczestnicy poprawnie rozpoznawali deepfake audio tylko w ~73% przypadków, a więc ich skuteczność nieznacznie przewyższała losowe zgadywanie. To pokazuje, że poleganie wyłącznie na ocenie człowieka jest ryzykowne, a automatyczne systemy detekcji audio są niezbędne (choć i one nie zawsze przewyższają intuicję człowieka, gdyż najlepsze modele dotąd osiągają porównywalny rząd skuteczności). Zaawansowane sieci neuronowe do detekcji *deepfake* (zarówno wideo, jak i audio)

[178] O. Wasiuta, S. Wasiuta, Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość, Annales Universitatis Paedagogicae Cracoviensis, Folia 289, Studia de Securitate, 9(3), 2019.

potrafią imponować wysoką skutecznością na zbiorach testowych, lecz ich wyniki w praktyce zależą od tego, na ile dane fałszerstwo przypomina te widziane podczas trenowania modelu. W warunkach laboratoryjnych odnotowano modele osiągające ponad 90% dokładności rozróżniania fałszywych twarzy od prawdziwych w znanych benchmarkach. W 2019 r. Facebook zorganizował nawet globalne zawody *Deepfake Detection Challenge*, aby zmobilizować naukowców do stworzenia najlepszych detektorów – zwycięski algorytm uzyskał około 82% skuteczności na udostępnionym publicznie zbiorze testowym. Jednak gdy sprawdzono go na nowym, ukrytym zestawie nagrań, skuteczność spadła do ~65%. Oznacza to, że prawdziwym sprawdzianem jest uogólnienie metody na nieznane wcześniej fałszywki. Najnowsze prace starają się poprawić tę generalizację – np. poprzez łączenie wielu modeli w jeden (*ensembling*) albo skupienie na cechach bardziej uniwersalnych. Przykładem jest zaproponowany model *Texture and Artifact Detector* (TAD), który jednocześnie analizuje teksturę skóry oraz typowe artefakty generowania, łącząc wyniki tych dwóch podejść dla lepszej odporności na nowy rodzaj manipulacji[179]. Innym kierunkiem jest poprawa detekcji z zakodowanych lub niskiej jakości materiałów – silna kompresja pliku wideo potrafi zamaskować drobne artefakty i znacząco obniżyć skuteczność wykrywania. Zaproponowano np. adaptacyjne filtry podbijające wysokie częstotliwości w obrazie wideo przed analizą, co ma wydobyć ukryte ślady manipulacji nawet w silnie skompresowanych nagraniach. Wszystko to pokazuje, że detekcja *deepfake'ów* jest wyścigiem zbrojeń między twórcami fałszywek a badaczami je demaskującymi.

Skuteczność obecnych metod w warunkach rzeczywistych pozostaje ograniczona. Szereg czynników sprawia, że nawet najlepszy wykrywacz może zawieść:

- Ograniczona generalizacja i podatność na obejście. Algorytmy wykrywające często są wyspecjalizowane pod kątem konkretnych właściwości fałszerstw. Gdy zmieni się technika generowania *deepfake'u* – np. pojawi się nowy model oparty na innych założeniach – dotychczasowe

[179] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, T.S. Bonaventura, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera, S. Mandelli, G.L. Marcialis, M. Micheletto, A. Montibeller, G. Orrù, A. Ortis, P. Perazzo, D. Salvi, S. Tubaro, C.M. Tonti, M. Villari, D. Vitulano, Deepfake Media Forensics: State of the Art and Challenges Ahead, arXiv preprint arXiv:2408.00388v1, 2024.

detektory mogą nie rozpoznać zagrożenia. Jak wskazuje raport GAO[180], skuteczność istniejących modeli znacząco spada, jeśli np. warunki oświetlenia, mimika twarzy czy jakość nagrania różnią się od danych treningowych, lub gdy użyto innej architektury generującej niż te, na których trenowano detektor. Co więcej, twórcy *deepfake'ów* aktywnie uczą się omijać znane metody wykrywania – pojawiają się coraz bardziej dopracowane fałszywki pozbawione dotychczasowych „oznak" fałszu (np. modele generujące nauczyły się już poprawnie odwzorowywać naturalne mruganie czy cienie). Istnieje też ryzyko celowych ataków na detektory: poprzez dodanie do fałszywego obrazu niewidocznych dla oka perturbacji (tzw. ataki *adversarialne*) można zmylić algorytm detekcyjny tak, by zaklasyfikował *deepfake* jako prawdziwy. Badacze bezpieczeństwa zwracają uwagę, że każda metoda detekcji może zostać prędzej czy później obejścia – jest to typowa dynamika broni i tarczy w cyberbezpieczeństwie. Przykładowo, gdy opracowano detektory skupiające się na nienaturalnym ruchu gałek ocznych, generatory szybko uwzględniły ten czynnik. Analogicznie, przyszłe postępy w algorytmach generujących mogą wyeliminować inne obecnie wykorzystywane *hallmarks* (jak choćby drobne przebarwienia skóry czy nienaturalne odbicia światła).

- Wyzwania techniczne i wydajnościowe. Skuteczna detekcja wymaga często intensywnej analizy każdego kadru wideo lub każdej próbki audio, co pociąga za sobą wysokie koszty obliczeniowe. Uruchomienie głębokiej sieci neuronowej na godzinnym materiale wideo w rozdzielczości HD może trwać bardzo długo, co utrudnia zastosowanie tych metod w czasie rzeczywistym. We wspomnianym konkursie Facebooka większość modeli była zbyt wolna, by użyć ich na żywo – skupiano się na maksymalizacji trafności, nie szybkości. W praktyce czas ma znaczenie, zwłaszcza gdy fałszywe treści rozprzestrzeniają się *viralowo* w mediach społecznościowych. Jeśli detekcja zajmuje godziny lub dni, może być spóźniona wobec cyklu informacyjnego. Ponadto, skuteczność detektorów bywa ograniczona przy gorszej jakości danych. Silna kompresja stratna (np. wideo przesłane przez platformy internetowe) usuwa wiele detali, które mogłyby zdradzić

[180] U.S. Government Accountability Office (GAO), Science & Tech Spotlight: Combating Deepfakes, GAO-24-107292, Waszyngton, 2024.

fałszerstwo – jak wskazują badania, dokładność wykrywania *deepfake* spada dramatycznie dla materiałów niskiej rozdzielczości lub z artefaktami kompresji. Detektor może wówczas generować zarówno fałszywe alarmy (gdy niska jakość prawdziwego nagrania zostanie błędnie zinterpretowana jako ślad manipulacji), jak i *false negative* – nie wychwycić faktycznie podmienionej twarzy, bo ślady zostały starte przez kompresję.

- Kontekst i ograniczenia pozatechnologiczne. Same algorytmy nie zawsze wystarczą – liczy się też kontekst zastosowania. Detektor działający skutecznie w kontrolowanych warunkach może zawodzić, gdy zostanie zastosowany do moderacji treści w Internecie na masową skalę. W praktyce system musi odróżniać *deepfake* od innych form fałszerstw (np. proste montaże, tzw. *cheapfakes*) oraz od prawdziwych nagrań. Błędne oznaczenie autentycznego materiału jako fałszywego stanowi ryzyko (technologia nie jest nieomylna – „*false positive*" może podważyć wiarygodność prawdziwej informacji). Dlatego często potrzebna jest weryfikacja przez człowieka lub dodatkowe źródła. Uwarunkowania kontekstowe obejmują też dostępność detekcji – nie wszystkie media przechodzą kontrolę, zwłaszcza gdy fałszywe nagranie pojawi się nagle na niszowej platformie czy w zamkniętych grupach komunikatora. Zanim dotrze do profesjonalnych weryfikatorów, może osiągnąć miliony odsłon. Wreszcie, kontekst kulturowy i językowy też ma znaczenie: modele uczone na anglojęzycznych *deepfake-'ach* mogą nie radzić sobie ze sfałszowanymi nagraniami np. w języku chińskim czy polskim, gdzie inne cechy mogą zdradzać falsyfikat.
- Czynnik ludzki i poznawczy. Istotnym ograniczeniem jest też percepcja odbiorców oraz zjawiska psychologiczne towarzyszące erze *deepfake*. Paradoksalnie, nawet doskonałe narzędzia detekcji nie gwarantują, że społeczeństwo będzie skutecznie chronione przed skutkami fałszywych treści. Po pierwsze, wiele osób może nie mieć dostępu do zaawansowanych detektorów lub nie umieć z nich korzystać – wówczas są zdani na własną ocenę, a ta, jak wspomniano, bywa zawodna. Po drugie, istnieje ryzyko tzw. dywidendy kłamcy (ang. *liar's dividend*), czyli sytuacji, w której samo istnienie technologii *deepfake* podważa zaufanie do autentycznych materiałów. Polityk przyłapany na kompromitującym nagraniu może twierdzić, że to *deepfake*

– i część odbiorców mu uwierzy, bo nauczyli się, że „dziś wszystko można sfałszować". Zjawisko to zaobserwowano już w praktyce: w 2020 r. w trakcie kampanii w USA pojawiały się bezpodstawne oskarżenia, że niewygodne materiały to fałszywki, co zaciemniało obraz debaty publicznej. Ponadto, nawet gdy *deepfake* zostanie wykryty i zdementowany, nie gwarantuje to odwrócenia jego wpływu. Fałszywe treści rozchodzą się często szybciej i szerzej niż sprostowania – odbiorcy mogą zapamiętać samą sensację, a sprostowania nie dotrzeć do wszystkich zainteresowanych. Jak obrazowo stwierdza raport GAO[181]: identyfikacja *deepfake'u* nie powstrzyma automatycznie szkód, bo dezinformacja może rozpowszechnić się zanim zostanie zdemaskowana. W efekcie społeczeństwo może popaść w stan ogólnej niepewności co do wiarygodności materiałów w mediach – jest to element szerszego problemu, jakim jest erozja zaufania społecznego w dobie tzw. post-prawdy.

Analiza dotychczasowych incydentów dowodzi, iż technologia *deepfake* odegrała istotną rolę w szeregu kluczowych zdarzeń z ostatnich lat, co potwierdza jej znaczenie dla szeroko pojętego bezpieczeństwa. Jednym z najbardziej nagłośnionych przypadków – o wysokiej doniosłości politycznej – było spreparowanie materiału wideo rzekomo przedstawiającego prezydenta Ukrainy w trakcie trwającego konfliktu zbrojnego. W marcu 2022 r., w szczytowym momencie rosyjskiej inwazji, w sieci pojawiło się nagranie sugerujące, że Wołodymyr Zełenski wzywa ukraińskich żołnierzy do złożenia broni i poddania się siłom rosyjskim. Materiał ten został szybko zidentyfikowany jako *deepfake*: stosunkowo mało zaawansowany technicznie, co zdradzały m.in. niespójny koloryt skóry twarzy i szyi, nienaturalna artykulacja oraz zauważalne artefakty wokół głowy[182]. Mimo iż film został przez krótki czas wyemitowany na zhakowanej ukraińskiej stronie telewizyjnej, oficjalne czynniki rządowe zdementowały jego treść. Sam prezydent Zełenski w trybie natychmiastowym opublikował wiarygodne oświadczenie, nazywając spreparowany klip „dziecinną prowokacją" i podkreślając, że nigdy nie wzywał do kapitulacji. Największe platformy (Facebook, YouTube) również niezwłocznie usunęły

[181] U.S. Government Accountability Office (GAO), Science & Tech Spotlight: Combating Deepfakes, GAO-24-107292, Waszyngton, 2024.

[182] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107(6), 1753-1820.

fałszywe nagranie. Incydent ten, choć nie odniósł realnego sukcesu dezinformacyjnego, został przez ekspertów uznany za zapowiedź bardziej zaawansowanych operacji informacyjnych w przyszłości[183]. Był zarazem pierwszym oficjalnie udokumentowanym przykładem intencjonalnego użycia *deepfake* w aktywnej wojnie informacyjnej: agresor (Rosja) podjął próbę osłabienia morale ukraińskich obrońców i wywołania chaosu w przestrzeni medialnej. Kijów wcześniej ostrzegał, iż tego rodzaju manipulacje mogą stanowić element rosyjskich działań propagandowych. Pomimo niepowodzenia w zakresie osiągnięcia efektu operacyjnego, przypadek ten uwidocznił decydentom i opinii międzynarodowej, że *deepfake* to nie tylko ciekawostka technologiczna, lecz potencjalny instrument prowadzenia walki informacyjnej.

W dalszej kolejności odnotowano kolejne incydenty potwierdzające użyteczność *deepfake* w dezinformacji i wojnie informacyjnej. Przykładem jest seria fałszywych połączeń wideo z czerwca 2022 r., kiedy to osobie podszywającej się pod mera Kijowa, Witalija Kłyczkę, udało się przeprowadzić rozmowy z przedstawicielami władz Berlina, Wiednia i Madrytu. Fałszywa postać wyświetlana na ekranie wiarygodnie imitowała cechy wizualne oraz głos Kłyczki, dzięki czemu politycy przez dłuższy czas nie dostrzegali manipulacji. Dopiero gdy podczas telekonferencji „Kłyczko" zaczął artykułować nietypowe postulaty (np. odesłanie uchodźców do Ukrainy w celu przymusowego wcielenia do wojska), strona niemiecka nabrała podejrzeń i przerwała transmisję, po czym natychmiast potwierdzono u prawdziwego Kłyczki, że nie prowadził on żadnej wideorozmowy w tym czasie. Władze Berlina wydały komunikat wskazujący na zastosowanie technologii *deepfake*, z kolei eksperci ocenili, że wysoki realizm tej symulacji stanowi istotne ostrzeżenie przed zagrożeniami w czasie rzeczywistym[184]. Nie ustalono oficjalnie sprawców, jednak sugerowano, że mogły to być grupy prorosyjskie dążące do podważenia zaufania i skłócenia europejskich partnerów Ukrainy. Incydent ten pokazał, że nawet doświadczeni decydenci polityczni bywają narażeni na skuteczne oszustwo w formie *deepfake* – co burmistrz Berlina podsumowała stwierdzeniem o „wszechstronnych metodach prowadzenia wojny, także w cyberprzestrzeni, w celu zniszczenia zaufania". Strategicznie

[183] Zob.Y. Mirsky, W. Lee, The Creation and Detection of Deepfakes: A Survey, ACM Computing Surveys, 54(1), 2021.

[184] Zob. R. Chesney, D. Citron, Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics, Foreign Affairs, 98(1), 147-155.

patrząc, takim fałszerstwem można było wywołać nieporozumienia między władzami miejskimi a ukraińskim sojusznikiem, co w konsekwencji miało prowadzić do podziałów politycznych. Na szczęście oszustwo wykryto stosunkowo szybko, lecz incydent wzmocnił sygnały ostrzegawcze dla unijnych i natowskich struktur odpowiedzialnych za bezpieczeństwo informacyjne.

Deepfake bywa również używany w kontekście cyberprzestępczym do celów finansowych. Ilustruje to głośna sprawa z 2019 r., kiedy to do impersonacji wykorzystano klon głosu (audio *deepfake*), aby podszyć się pod prezesa niemieckiego koncernu energetycznego wobec dyrektora finansowego brytyjskiej spółki zależnej. Oszuści przekonali rozmówcę do przelania 243 tys. dolarów, argumentując pilną i poufną transakcją. W toku późniejszego dochodzenia ustalono, że do wiernego odwzorowania akcentu oraz charakterystycznych modulacji głosu posłużono się oprogramowaniem AI. Za pierwsze udokumentowane oszustwo finansowe typu „*deepfake audio*" uznaje się właśnie ten incydent. Mechanizm opierał się na przełamaniu naturalnego sposobu weryfikacji tożsamości (rozpoznanie głosu przełożonego). W latach kolejnych odnotowano próby powtórzenia takiego schematu w korporacjach o globalnym zasięgu, a w 2023 r. także Federalne Biuro Śledcze (FBI) wydało ostrzeżenie o rosnącej liczbie przypadków wyłudzeń z użyciem klonowania głosów osób bliskich proszących o nagłą pomoc finansową[185]. Choć nie są to akty o charakterze stricte militarnym, szkody wynikające z tej formy oszustwa dowodzą, że deepfake stanowi zagrożenie nie tylko dla bezpieczeństwa politycznego, lecz również ekonomicznego.

Wskazane przykłady obrazują szeroki zakres możliwych konsekwencji *deepfake* w wymiarze bezpieczeństwa narodowego, stabilności demokratycznej i ładu społecznego. W kontekście politycznym i informacyjnym *deepfake* może prowadzić do destabilizacji systemu wyborczego, na przykład poprzez nagłośnienie spreparowanego nagrania kandydata na urząd polityczny, wygłaszającego skandaliczne lub wrogie społeczeństwu tezy tuż przed dniem głosowania[186]. Nawet jeśli materiał szybko okaże się fałszywy, jego wpływ na preferencje wyborców może być nieodwracalny. Co więcej, rosnąca świadomość możliwości fabrykowania treści sprawia, że politycy

[185] FBI Warns of Voice Cloning Scams Using AI-Generated Audio, Internet Crime Complaint Center Public Service Announcement, 2023.

[186] Zob. R. Chesney, D. Citron, Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics, Foreign Affairs, 98(1), s. 147-155.

oskarżani o nadużycia mogą podważać autentyczność prawdziwych dowodów, nazywając je „*deepfake*". To z kolei degraduje zaufanie do instytucji i mediów, promując postawy oparte na spiskowych teoriach i polaryzacji społecznej. Jak zauważają badacze, jeśli społeczeństwo utraci pewność co do wiarygodności nawet podstawowych materiałów wideo i audio, może dojść do głębokiego kryzysu „wspólnej płaszczyzny prawdy".[187]

Nie mniejsze zagrożenia dotyczą obronności państwa i systemów bezpieczeństwa. Fałszywy rozkaz wydany rzekomo przez dowódcę wojskowego może wywołać dezorientację w strukturach siłowych, a spreparowana wypowiedź przywódcy państwowego – zawierająca obelgi w stronę innych krajów – potencjalnie wywołać napięcia dyplomatyczne. Już teraz w NATO i Unii Europejskiej prowadzi się analizy scenariuszy, w których przeciwnik fabrykuje materiały wizyjne dokumentujące rzekomy atak lub zbrodnię wojenną, aby uzasadnić retorsje lub sprowokować rozłam w sojuszu. Z perspektywy społecznej długofalowym efektem może być dalsza erozja zaufania publicznego: w świecie, w którym każdą treść można uznać za „*fejk*", obywatele popadają w cynizm informacyjny i odwracają się od tradycyjnych źródeł informacji[188]. Taka „amnezja faktów" sprzyja rozprzestrzenianiu się manipulacji i retoryk autorytarnych, opartych na podważaniu fundamentów racjonalnego dyskursu.

W reakcji na powyższe zagrożenia coraz częściej akcentuje się wielotorowe strategie obrony, nieograniczające się wyłącznie do detekcji fałszerstw post factum. Pierwszym filarem jest rozwijanie aktywnych metod uwierzytelniania treści. Zamiast koncentrować się jedynie na wychwytywaniu śladów manipulacji, proponuje się wprowadzenie mechanizmów, które od początku pozwolą odróżnić materiał oryginalny od zmanipulowanego. Jednym z rozwiązań jest cyfrowe znakowanie danych (tzw. *watermarking*), polegające na umieszczaniu w zdjęciach lub nagraniach wzorców niewidocznych dla ludzkiego oka, ale wykrywalnych przez oprogramowanie. Najmniejsza ingerencja w plik powoduje zniekształcenie znaczników, co w procesie weryfikacji ujawnia manipulację. Alternatywnie można stosować kryptograficznie zabezpieczone metadane z informacją o pochodzeniu

[187] Zob. Y. Mirsky, W. Lee, The Creation and Detection of Deepfakes: A Survey, ACM Computing Surveys, 54(1), 2021.

[188] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107(6), s. 1753-1820.

i historii edycji materiału. Takie podejście jest rozwijane m.in. w ramach *Content Authenticity Initiative* (CAI) i standaryzacji C2PA (ang. *Coalition for Content Provenance and Authenticity*), w której uczestniczą czołowe platformy internetowe, agencje prasowe oraz producenci sprzętu foto-wideo. W 2023 r. zaczęły się także pojawiać regulacje prawne na poziomie Unii Europejskiej (Akt o Sztucznej Inteligencji) i w Chinach, nakazujące jawne oznaczanie materiałów syntetycznych. Takie działania tworzą infrastrukturę zaufania w mediach cyfrowych i mają utrudniać fałszerzom anonimowe publikowanie spreparowanych treści.

Drugim fundamentem jest edukacja informacyjna i budowanie świadomości społecznej. Z uwagi na fakt, że żadna pojedyncza technologia nie gwarantuje stuprocentowej skuteczności, konieczne jest kształtowanie postaw krytycznego odbioru treści medialnych. Organizacje międzynarodowe np. UNESCO[189] i *fact-checkerskie* rozwijają programy szkoleniowe, uczące identyfikowania znaków możliwej manipulacji (nietypowa mimika, zniekształcenia w krawędziach obrazu, nienaturalny ruch ust, dziwne przejścia w modulacji głosu). W kontekście wyborczym stosuje się tzw. *prebunking* – uprzedzanie społeczeństwa o możliwości pojawienia się głęboko spreparowanych materiałów w kampaniach politycznych. Dzięki temu wyborcy czy politycy, konfrontując się z zaskakującym nagraniem, będą bardziej skłonni poddać je analizie i weryfikacji w wiarygodnym źródle. Analogiczne szkolenia prowadzi także NATO, przygotowując żołnierzy i oficerów do rozpoznawania fałszywych rozkazów czy medialnych prowokacji. Tak zwane „kompetencje medialne 2.0" obejmują więc nie tylko umiejętność analizy tekstu, ale również rozumienie potencjalnych manipulacji w warstwie audio-wideo generowanych przez AI.

W syntetycznym ujęciu, problem wykrywania *deepfake* to złożone zagadnienie technologiczne i społeczne, wymagające podejścia wielowarstwowego. Aktualnie dostępne rozwiązania oparte na zaawansowanych algorytmach uczenia maszynowego pozwalają na skuteczne wyłapywanie wielu rodzajów manipulacji, jednak w warunkach realnych – cechujących się

[189] UNESCO (ang. United Nations Educational, Scientific and Cultural Organization), Organizacja Narodów Zjednoczonych ds. Edukacji, Nauki i Kultury), agencja ONZ założona w 1945 roku, z siedzibą w Paryżu, działająca na rzecz promowania pokoju, bezpieczeństwa i współpracy międzynarodowej poprzez wspieranie edukacji, nauki, kultury oraz komunikacji.

różnorodnością źródeł i szybką ewolucją narzędzi generujących – ich działanie ulega ograniczeniom. Stąd konieczne są działania na polu technicznym (narzędzia detekcji, systemy cyfrowego znakowania, standaryzacja metadanych), organizacyjnym (międzynarodowe regulacje i współpraca w ramach sojuszy) oraz społecznym (edukacja medialna i kampanie uświadamiające). Tylko w taki sposób można przeciwdziałać ryzyku, iż *deepfake* doprowadzi do destabilizacji ładu społeczno-politycznego, zagrozi bezpieczeństwu narodowemu czy osłabi instytucje publiczne opierające się na zaufaniu i rzetelnej informacji. Społeczność międzynarodowa – od laboratoriów akademickich po agencje rządowe – podejmuje intensywne wysiłki w tym zakresie, świadoma, że stawką jest utrzymanie wiarygodności fundamentów życia publicznego w dobie coraz doskonalszych „syntetycznych rzeczywistości".

5.5. Uwarunkowania prawne i stanowiska międzynarodowe w kwestii monitorowania treści

Rozwój technologii analizy treści audiowizualnych stawia pytania o granice prawne ich wykorzystywania. Poszczególne państwa przyjmują odmienne strategie równoważenia bezpieczeństwa informacyjnego z ochroną praw obywatelskich (wolności słowa, prywatności). W Europie kładzie się nacisk na regulacje zwiększające odpowiedzialność platform internetowych za przeciwdziałanie dezinformacji – przykładem jest niemiecka ustawa NetzDG zobowiązująca serwisy społecznościowe do szybkiego usuwania nielegalnych treści, czy francuska ustawa przeciw manipulacjom informacją z 2018 r., przewidująca m.in. możliwość sądowego zablokowania oczywiście fałszywych informacji podczas kampanii wyborczych. Równolegle Unia Europejska wypracowuje pierwsze kompleksowe prawo dotyczące AI – Akt o Sztucznej Inteligencji. Choć nie zakazuje on *deepfake'ów* per se, nakłada surowe wymogi transparentności na podmioty tworzące i rozpowszechniające syntetyczne media[190]. Zgodnie z projektowanymi przepisami UE każdy materiał wygenerowany przy udziale AI (w szczególności imitujący rzeczywiste osoby lub zdarzenia) będzie musiał być wyraźnie oznaczony jako sztuczny, a twórcy takich treści będą zobowiązani do prowadzenia dokumentacji

[190] European Commission, Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), European Commission, 2021.

wyjaśniającej, jak dany *deepfake* został wytworzony. Dodatkowo rozważa się ograniczenia użycia *deepfake'ów* w niektórych kontekstach wysokiego ryzyka – np. w działaniach mogących naruszać porządek publiczny lub prowadzić do masowej dezinformacji. Już teraz w projekcie AI Act sugeruje się zakaz wykorzystywania *deepfake'ów* do nielegalnej inwigilacji czy innych działań zagrażających bezpieczeństwu państwa. Ochronie obywateli służą też przepisy o danych osobowych: w UE wizerunek osoby jest prawnie chronioną daną, więc wykorzystanie czyjegoś wizerunku w *deepfake'u* bez zgody może naruszać RODO. Podobnie w Wielkiej Brytanii byłoby to sprzeczne z Data Protection Act (2018). W Stanach Zjednoczonych brak jest federalnej ustawy zakazującej dezinformacji czy używania *deepfake'ów* – przede wszystkim ze względu na silne gwarancje Pierwszej Poprawki dotyczące wolności słowa[191]. Mimo to wprowadzono pewne ograniczone przepisy na poziomie stanowym. Przykładowo, Kalifornia penalizuje rozpowszechnianie fałszywych nagrań wizerunkowych polityków na krótko przed wyborami, a Wirginia zakazała tzw. pornografii *deepfake* (niezgodnego z wolą wykorzystania czyjegoś wizerunku w treściach erotycznych). Ponadto kilka stanów (m.in. Kolorado) przyjęło ramy prawne regulujące rozwój systemów AI pod kątem etycznym, gdzie zagadnienie deepfake również się pojawia (np. obowiązek oznaczania materiałów syntetycznych w sektorze publicznym – Yampolskiy, 2021). Chiny i Rosja stawiają na restrykcyjną kontrolę obiegu treści. W Chinach od początku 2023 r. obowiązuje prawo wymagające oznaczania wszystkich syntetycznych obrazów i nagrań oraz zakazujące publikacji *deepfake'ów* bez wyraźnej informacji o ich sztucznym charakterze[192]. Regulacje te – określane jako przepisy o tzw. "głębokiej syntezie" – są częścią szerszych działań Chińskiej Administracji Cyberprzestrzeni, mających na celu zapobieganie „zagrożeniom dla stabilności społecznej" wynikającym z niekontrolowanego użycia AI. Z kolei Rosja w ostatnich latach wprowadziła przepisy wymierzone w zwalczanie dezinformacji w przestrzeni publicznej – m.in. ustawę z 2019 r. zakazującą rozpowszechniania „fałszywych informacji" pod groźbą kar, a w 2022 r. zaostrzyła kodeks karny w zakresie kar za „umyślne rozpowszechnianie nieprawdziwych informacji" o działaniach sił zbrojnych.

[191] Zob. R. Chesney, D. Citron, Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics, Foreign Affairs, 98(1), 147-155.

[192] Zob. D. Laje, China's Deep Fake Law Is Fake, Signal Magazine (AFCEA International), 2023.

W obu tych państwach monitorowanie i cenzurowanie treści jest elementem polityki tzw. bezpieczeństwa ideologicznego, co rodzi obawy o nadużycia – prawo bywa wykorzystywane jako narzędzie tłumienia niezależnych źródeł informacji i opozycji. Organizacje międzynarodowe podkreślają potrzebę współpracy i wypracowania wspólnych standardów wobec omawianego problemu. ONZ w kontekście tzw. *infodemii* (zalewu dezinformacji podczas pandemii COVID-19) apelowała o globalną koordynację działań i wymianę informacji między państwami. Jednocześnie raporty UNESCO wskazują, że walka z dezinformacją nie może odbywać się kosztem wolności słowa – potrzebne jest wyważone podejście chroniące prawa człowieka[193]. Proponuje się rozwiązania oparte na zwiększaniu transparentności algorytmów platform, edukacji medialnej społeczeństw oraz dobrowolnych kodeksach postępowania dla firm technologicznych (takich jak Kodeks postępowania UE w zakresie dezinformacji z 2018 r.). NATO od kilku lat traktuje dezinformację jako element wojny hybrydowej zagrażający bezpieczeństwu Sojuszu. W 2024 r. państwa NATO przyjęły wspólną strategię przeciwdziałania zagrożeniom informacyjnym, akcentując m.in. ryzyko związane z deepfake'ami używanymi przez aktorów wrogich do podważania zaufania społecznego. NATO i UE zacieśniają również współpracę w tym obszarze, aby wymieniać się danymi i technologiami do wykrywania fałszerstw na masową skalę. Eksperci zauważają, że harmonizacja regulacji międzynarodowych będzie konieczna, by skutecznie ograniczyć negatywne skutki *deepfake'ów* – bez wspólnych standardów twórcy fałszywek mogą bowiem przenosić się do jurysdykcji o luźniejszych przepisach[194]. Podsumowując, ramy prawne dotyczące monitorowania i przeciwdziałania manipulacji treściami w mediach wciąż ewoluują: od pierwszych ustaw *anty–fake news* w poszczególnych krajach, poprzez unijne pionierskie regulacje w zakresie sztucznej inteligencji, aż po globalne dyskusje w ONZ i na forum G7 nad tym, jak pogodzić bezpieczeństwo informacyjne z poszanowaniem wolności obywatelskich[195]. Poniższe zestawienia w tabeli 2 i tabeli 3 przedstawiają zróżnicowane podejścia prawne wybranych państw oraz inicjatywy międzynarodowe:

[193] Zob. UNESCO, Threat of Disinformation – Report of the UNESCO International High-Level Panel on Information Democracy, Paris: UNESCO, 2021.

[194] Zob. M. Westerlund, The Emergence of Deepfake Technology: A Review, Technology Innovation Management Review, 9(11), 39-52.

[195] Zob. G7, Hiroshima Leaders' Communiqué: G7 Summit, (Ai governance and disinformation highlighted), 2023.

Tabela 2. Przykłady regulacji krajowych dotyczących monitorowania treści audiowizualnych

Jurysdykcja	Podejście prawne do zmanipulowanych treści
Europa (UE)	Odpowiedzialność platform za treści: m.in. niemiecka ustawa NetzDG (2017) wymusza szybkie usuwanie nielegalnych materiałów, a francuska ustawa (2018) przeciw manipulacji informacją umożliwia blokowanie *fake news* w trakcie wyborów. Trwają prace nad *Aktem o SI*, który nałoży obowiązek wyraźnego oznaczania treści generowanych przez AI i rozważa zakaz użycia *deepfake* w niektórych kontekstach. Ochronę praw obywateli wzmacnia RODO (wizerunek jako dana osobowa).
USA	Brak ogólnokrajowego zakazu *deepfake* (Pierwsza Poprawka chroni wolność słowa). Wprowadzono jednak przepisy stanowe: np. Kalifornia zakazuje fałszywych materiałów o kandydatach na krótko przed wyborami, Wirginia penalizuje pornograficzne *deepfake'i*. Niektóre stany (np. Kolorado) przyjmują także wytyczne etyczne dla AI, uwzględniające oznaczanie treści syntetycznych.
Chiny	Surowe regulacje (od 2023) wymagają jednoznacznego oznaczania wszelkich syntetycznych obrazów i nagrań; publikacja *deepfake* bez takiego oznaczenia jest zakazana. Monitorowanie treści jest elementem państwowej cenzury i polityki „bezpieczeństwa ideologicznego".
Rosja	Restrykcyjna kontrola informacji: prawo zakazuje „*fake news*" (ustawa 2019) i dezinformacji szczególnie o siłach zbrojnych (2022), przewidując kary więzienia. Władze intensywnie monitorują media i Internet, oficjalnie dla ochrony stabilności państwa – co bywa wykorzystywane do tłumienia niezależnych przekazów.

Źródło: opracowanie własne.

Tabela 3. Stanowiska wybranych organizacji międzynarodowych w sprawie dezinformacji i *deepfake*

Podmiot	Inicjatywy i stanowiska
ONZ	Wezwania Sekretarza Generalnego do globalnej koordynacji walki z dezinformacją (szczególnie podczas pandemii COVID-19). Podkreślanie konieczności wypracowania standardów przeciwdziałania manipulacjom informacyjnym przy jednoczesnym poszanowaniu praw człowieka i wolności słowa.
UNESCO	Raporty (2020–2022) przestrzegające, że walka z *fake news* nie może odbywać się kosztem wolności mediów. Rekomendacje obejmują zwiększenie transparentności algorytmów, edukację medialną społeczeństw oraz dobrowolne kodeksy postępowania dla platform (np. unijny Kodeks postępowania ws. dezinformacji z 2018 r.).
NATO	Uznanie dezinformacji (w tym *deepfake*) za zagrożenie hybrydowe dla bezpieczeństwa Sojuszu. Przyjęcie w 2024 r. strategii przeciwdziałania zagrożeniom informacyjnym, akcentującej potrzebę monitorowania i szybkiego demaskowania fałszywych treści. Współpraca z UE nad wymianą danych i technologii wykrywania manipulacji (m.in. w ramach NATO *StratCom COE*).
G7	Wspólne deklaracje państw G7 wzywające do ustanowienia międzynarodowych „*guardrails*" (barier ochronnych) dla sztucznej inteligencji generatywnej. W ramach *Rapid Response Mechanism* (RRM), utworzonego w 2018 r., kraje G7 wymieniają się informacjami o kampaniach dezinformacyjnych i koordynują działania przeciw złośliwym aktorom ingerującym w procesy demokratyczne.

Źródło: opracowanie własne.

VI

Zastosowanie algorytmów analizy treści audiowizualnych w bezpieczeństwie narodowym

6.1. Algorytmy rozpoznawania obiektów audiowizualnych

Rozpoznawanie obiektów audiowizualnych stanowi obecnie jeden z filarów nowoczesnych systemów bezpieczeństwa państwa. W warunkach intensyfikacji zagrożeń hybrydowych, obejmujących m.in. masową cyfryzację przekazów informacyjnych, rozwój sieci telewizji dozorowej CCTV (ang. *Closed-Circuit Television*) i powszechne udostępnianie treści w mediach społecznościowych, konieczne stało się wdrażanie zaawansowanych narzędzi umożliwiających automatyczną analizę obrazów i dźwięków. Z perspektywy nauk o bezpieczeństwie istotne jest zrozumienie, jak narzędzia te wpływają na skuteczność działań prewencyjnych, operacyjnych oraz dowodowych, zwłaszcza w kontekście ochrony przestrzeni publicznej i reagowania na nowe formy przestępczości oraz destabilizacji informacyjnej.

Początkowy rozwój metod analizy treści audiowizualnych opierał się na tzw. klasycznych algorytmach, polegających na ręcznym definiowaniu cech obrazu lub sygnału dźwiękowego. W zakresie wizji komputerowej stosowano m.in. histogramy zorientowanych gradientów, pozwalające na identyfikację lokalnych konturów i kształtów obiektów[196,197], czy też klasyfikator

[196] Zob. M. Chodyka, Algorytmy rozpoznawania logo oraz oznaczeń kategorii wiekowej przekazu telewizyjnego w sekwencji obrazów, Rozprawa doktorska, Politechnika Łódzka, 2012.

[197] Zob. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, 2005, 886-893.

Viola-Jones oparty na cechach Haar i metodzie AdaBoost, przeznaczony do wykrywania twarzy w obrazach[198]. W dziedzinie rozpoznawania mowy oraz zdarzeń akustycznych dominowały parametry MFCC (ang. *Mel-Frequency Cepstral Coefficients*) oraz modele ukrytych modeli Markowa HMM, (ang. *Hidden Markov Model*), wykorzystywane do przetwarzania sekwencji fonemów[199]. Pomimo swojej przydatności, podejścia te charakteryzowały się niską elastycznością i ograniczoną zdolnością uogólniania w zmiennych warunkach środowiskowych. Nie były również dostosowane do przetwarzania danych w dużej skali, co ograniczało ich zastosowanie w kontekstach takich jak analiza wielogodzinnych nagrań z miejskiego monitoringu czy detekcja złożonych wzorców przestępczych. Punktem zwrotnym w rozwoju technologii rozpoznawania obiektów stało się wprowadzenie metod głębokiego uczenia (ang. *deep learning*), opartych na wielowarstwowych sieciach neuronowych[200].
Szczególnie istotną rolę w dziedzinie przetwarzania obrazu odegrały konwolucyjne sieci neuronowe CNN (ang. *Convolutional Neural Network*), które dzięki możliwości automatycznego uczenia się hierarchii cech bez konieczności ich wcześniejszego definiowania przez eksperta, znacząco poprawiły skuteczność klasyfikacji i detekcji obiektów. Przykładem zastosowania tych architektur była sieć *AlexNet*, która w 2012 roku zrewolucjonizowała konkurs *ImageNet*, wyznaczając nowe standardy w zakresie rozpoznawania obrazów[201]. Rozwój sieci CNN doprowadził do powstania modeli takich jak YOLO (ang. *You Only Look Once*), cechujących się zdolnością do jednoczesnej detekcji i klasyfikacji obiektów w czasie rzeczywistym[202], a także modeli dwustopniowych, np. *Faster* R-CNN, łączących wysoką precyzję z możliwością wykorzystania w analizie materiałów dowodowych[203]. Przykładowy

[198] Zob. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, 2001, 511-518.

[199] Zob. L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77(2), 1989, 257-286.

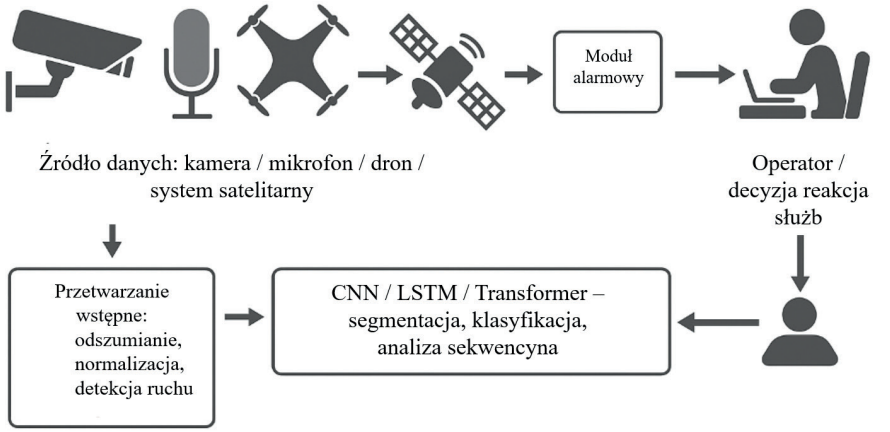
[200] Zob. Y. LeCun, Y. Bengio, G. Hinton, *Deep learning*, Nature, 521(7553), 436-444.

[201] Zob. A. Krizhevsky, I. Sutskever, G.E. Hinton, *Imagenet classification with deep convolutional neural network,* Advances in neural information processing systems, 25, 2012.

[202] Zob. J. Redmon, A. Farhadi, *YOLOv3: an incremental improvement*, arXiv preprint arXiv:1804.02767, 2018, s. 1-6.

[203] Zob. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems, 28, 91-99.

schemat architektury systemu analizy wideo w czasie rzeczywistym na potrzeby bezpieczeństwa operacyjnego zaprezentowany został na rysunku 3.



Rysunek 3. Schemat architektury systemu analizy wideo w czasie rzeczywistym na potrzeby bezpieczeństwa operacyjnego

Źródło: opracowanie własne.

W przedstawionym schemacie ukazano architekturę systemu analizy wideo w czasie rzeczywistym, zaprojektowanego z myślą o zastosowaniach w obszarze bezpieczeństwa operacyjnego. Proces rozpoczyna się od pozyskiwania danych z różnych źródeł: kamer monitoringu, mikrofonów, dronów oraz systemów satelitarnych. Dane te trafiają następnie do modułu przetwarzania wstępnego, w którym poddawane są operacjom takim jak odszumianie, normalizacja czy detekcja ruchu – co pozwala na ich przygotowanie do dalszej analizy. Kolejnym etapem jest analiza właściwa, realizowana przez wyspecjalizowane architektury sztucznej inteligencji, takie jak konwolucyjne sieci neuronowe (CNN), rekurencyjne sieci LSTM (ang. *Long Short-Term Memory*) lub nowoczesne modele typu transformer. W tym segmencie system dokonuje segmentacji obrazu, klasyfikacji obiektów oraz analizy sekwencji czasowych, co umożliwia wykrycie istotnych zdarzeń, zachowań lub obiektów o znaczeniu operacyjnym. Wyniki analizy trafiają do modułu alarmowego lub bazy danych, gdzie mogą być przechowywane, agregowane lub przetwarzane dalej, a także służyć jako podstawa do wygenerowania alertu. Ostateczne decyzje podejmowane są przez operatora systemu lub automatycznie przekazywane do odpowiednich służb, które mogą

podjąć natychmiastowe działania w odpowiedzi na zagrożenie. Tego rodzaju systemy są coraz powszechniej wykorzystywane w monitoringu miejskim, ochronie infrastruktury krytycznej, a także w operacjach wywiadowczych i antyterrorystycznych.

Analogiczny postęp obserwowano w analizie sygnału dźwiękowego. Zastosowanie konwolucyjnych sieci neuronowych do przetwarzania spektrogramów akustycznych[204], rekurencyjnych sieci LSTM (ang. *Long Short--Term Memory*)[205], a następnie transformatorów znanych z przetwarzania języka naturalnego[206], umożliwiło automatyczne rozpoznawanie zdarzeń dźwiękowych oraz mowy z dokładnością przekraczającą możliwości klasycznych algorytmów. W praktyce służb bezpieczeństwa przekłada się to na zdolność szybkiej transkrypcji przechwyconych rozmów, wykrywania sygnałów zagrożeń (np. odgłosy broni) czy analizę foniczną nagrań z miejsc przestępstw[207]. Wdrażanie technologii opartych na głębokim uczeniu w instytucjach bezpieczeństwa przyniosło zauważalne korzyści praktyczne. Przykładowo, systemy miejskiego monitoringu wzbogacone o sieci neuronowe umożliwiają automatyczne wykrywanie sytuacji podejrzanych (np. pozostawionego bagażu, zachowań agresywnych), rozpoznawanie twarzy osób poszukiwanych w tłumie[208], a także wsparcie działań operacyjnych poprzez analizę nagrań pod kątem występowania określonych obiektów[209]. W ana-

---

[204] Zob. N. Takahashi, M. Gygli, B. Pfister, L. Van Gool, Deep convolutional neural networks and data augmentation for acoustic event detection, Interspeech 2016, 2982-2986.

[205] Zob. A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, 2013, s. 6645-6649.

[206] Zob. L. Dong, S. Xu, B. Xu, Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, 2018, 5889-5893.

[207] Zob. H. Fayyad-Kazan, A. Hejase, I. Moukadem, S. Kassem-Moussa, Verifying the Audio Evidence to Assist Forensic Investigation, Computer and Information Science, Canadian Center of Science and Education, 14(3), 1-25.

[208] Zob. C.C. Loy, D. Lin, W. Ouyang, Y. Xiong, S. Yang, Q. Huang, D. Zhou, W. Xia, Q. Li, P. Luo, J. Yan, J. Wang, Z. Li, Y. Yuan, B. Li, S. Shao, G. Yu, F. Wei, X. Ming, D. Chen, S. Zhang, C. Chi, Z. Lei, S.Z. Li, H. Zhang, B. Ma, H. Chang, S. Shan, X. Chen, W. Liu, B. Zhou, H. Li, P. Cheng, T. Mei, A. Kukharenko, A. Vasenin, N. Sergievskiy, H. Yang, L. Li, Q. Xu, Y. Hong, L. Chen, M. Sun, Y. Mao, S. Luo, Y. Li, R. Wang, Q. Xie, Z. Wu, L. Lu, Y. Liu, W. Zhou, WIDER Face and Pedestrian Challenge 2018: Methods and Results, arXiv preprint arXiv:1902.06854, 2019.

[209] Zob. M. Ying Yang, W. Liao, C. Yang, Y. Cao, B. Rosenhahn, Security Event Recognition for Visual Surveillance, arXiv preprint arXiv:1810.11348, 2018.

lizie kryminalistycznej modele rozpoznawania umożliwiają szybkie przeszukiwanie i filtrowanie wielogodzinnych materiałów wideo w celu odnalezienia np. pojazdów o konkretnej marce, fragmentów broni czy elementów ubioru. Technologia ta znacząco skraca czas pracy analityków i zwiększa efektywność procesów dowodowych. Szczególnym wyzwaniem ostatnich lat stało się rozpoznawanie sfałszowanych treści audiowizualnych, generowanych technikami *deepfake*, opartymi na generatywnych sieciach przeciwstawnych GAN (ang. *Generative Adversarial Network*)[210]. Fałszywe nagrania prezentujące spreparowane wypowiedzi osób publicznych lub fałszywe wydarzenia stanowią poważne zagrożenie dla bezpieczeństwa informacyjnego państwa, w tym dla integralności procesów wyborczych i zaufania społecznego[211]. W odpowiedzi na te zagrożenia rozwijane są algorytmy detekcji, takie jak *MesoNet*, bazujące na CNN analizujących mikroskopijne artefakty manipulacji w obrazach wideo[212], czy systemy z rodziny FaceForensics++, wykorzystujące głębokie sieci do rozpoznawania anomalii w mimice i synchronizacji ust z dźwiękiem[213]. Coraz częściej stosuje się podejście multimodalne, które łączy analizę obrazu, dźwięku i tekstu w celu zwiększenia skuteczności wykrywania zmanipulowanych treści[214].

Pomimo znaczących osiągnięć, zastosowania technologii głębokiego uczenia w rozpoznawaniu obiektów audiowizualnych napotykają liczne wyzwania. Wymagają one olbrzymich zbiorów danych do nauki, co w kontekście bezpieczeństwa oznacza konieczność przetwarzania wrażliwych

[210] Zob. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, Generative Adversarial Nets, Advances in Neural Information Processing Systems, 27, 2014, s. 2672-2680.

[211] Zob. F. Marra, D. Gragnaniello, L. Verdoliva, G. Poggi, Do GANs leave artificial fingerprints?, 2019 IEEE Conference on Multimedia Information Processing and Retrieval, San Jose, CA, 2019, 506-511.

[212] Zob. D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a Compact Facial Video Forgery Detection Network, 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 2018, 1-7.

[213] Zob. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, FaceForensics++: Learning to detect manipulated facial images, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, 2019, 1-11.

[214] Y.H.H. Tsai, S. Bai, L.P. Morency, Multimodal transformer for unaligned multimodal language sequences, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, 2019, s. 655-660.

informacji, często objętych tajemnicą służbową[215]. Dodatkowo systemy te są energochłonne i obliczeniowo kosztowne, co ogranicza ich wdrażanie w urządzeniach mobilnych czy systemach czasu rzeczywistego. Kolejnym problemem jest ograniczona interpretowalność modeli – decyzje podejmowane przez sieci neuronowe trudno wytłumaczyć, co stanowi przeszkodę w kontekście wymogów procesowych i transparentności[216]. Równie poważne są zagrożenia wynikające z tzw. ataków *adversarialnych*, w których niewielkie zmiany w obrazie lub dźwięku prowadzą do błędnej klasyfikacji przez model[217]. Z perspektywy nauk o bezpieczeństwie zastosowania technologii rozpoznawania obiektów audiowizualnych obejmują nie tylko klasyczny nadzór wizyjny, lecz również działania w przestrzeni informacyjnej, ochronę infrastruktury krytycznej, kontrolę granic, cyberbezpieczeństwo oraz predykcję zachowań społecznych. Coraz częściej integruje się systemy rozpoznawcze z platformami analizy big data i systemami eksperckimi, co pozwala na prowadzenie predykcyjnej analizy zagrożeń oraz operacji informacyjnych[218]. W obszarze praktyki administracyjnej i wojskowej obserwuje się również rosnące zainteresowanie modelem human-in-the-loop, zakładającym współpracę człowieka z systemem sztucznej inteligencji w celu wzajemnego uzupełniania kompetencji i redukcji błędów[219].

Podsumowując, ewolucja algorytmów rozpoznawania obiektów audiowizualnych – od podejść klasycznych po zaawansowane modele głębokiego uczenia – znacząco zwiększyła zdolność służb bezpieczeństwa do automatycznego wykrywania, śledzenia i klasyfikowania zdarzeń i obiektów w obrazach i sygnałach dźwiękowych. Choć technologie te wymagają odpowiedniego przygotowania infrastrukturalnego oraz uwzględnienia aspektów prawnych i etycznych, ich rola w nowoczesnym systemie bezpieczeństwa

---

[215] Zob. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, s. 770-778.

[216] Zob. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys, 51(5), s. 1-42.

[217] Zob. A. Kurakin, I. Goodfellow, S. Bengio, Adversarial machine learning at scale, 5th International Conference on Learning Representations (ICLR), Toulon, 2017, s. 1-17.

[218] Zob. F. Turchini, L. Seidenari, T. Uricchio, A. Del Bimbo, Deep Learning Based Surveillance System for Open Critical Areas, Inventions, 3(4), 2018, s. 69.

[219] Zob. W. Samek, T. Wiegand, K.R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, ITU Journal: ICT Discoveries, 1(1), 39-48.

narodowego jest niepodważalna. Wdrażanie, rozwój i regulacja narzędzi analizy audiowizualnej stanowią istotne wyzwanie interdyscyplinarne – łączące informatykę, inżynierię danych, psychologię poznawczą oraz prawo – którego skuteczne rozwiązanie będzie determinować zdolność państw do przeciwdziałania zagrożeniom w środowisku informacyjnym XXI wieku.

6.2. Porównanie klasycznych i nowoczesnych metod analizy wideo

Analiza wideo wywodzi się z szerzej rozumianej dziedziny wizji komputerowej (ang. *computer vision*), która od lat 90. ubiegłego wieku rozwija się niezwykle dynamicznie. Rozwój ten wynika zarówno z rosnących potrzeb technologicznych (np. w zakresie monitoringu czy automatycznej interpretacji zdarzeń w ruchomym obrazie), jak i z postępów w dziedzinie sztucznej inteligencji. Metody analizy wideo dzieli się na dwie zasadnicze grupy: klasyczne (oparte na ręcznie definiowanych cechach, tzw. *feature engineering*) i nowoczesne (bazujące na uczeniu głębokim). W kontekście systemów bezpieczeństwa państwa porównanie tych podejść pozwala lepiej zrozumieć, w jakim kierunku zmierzają obecne i przyszłe rozwiązania służące automatycznej interpretacji danych wizualnych. W klasycznych metodach analiza wideo polega głównie na projektowaniu zestawów cech przez inżynierów lub badaczy, którzy z góry ustalają, co będzie istotne w obrazie. Następnie, na podstawie tak opisanych cech, stosuje się wybrane algorytmy uczenia maszynowego, takie jak maszyna wektorów nośnych SVM, (ang. *Support Vector Machine*) czy metody drzew decyzyjnych[220]. Przykładem jest proste podejście do wykrywania ruchu w wideo: najpierw aktualizowany jest model tła (ang. *background subtraction*), a wszelkie odstępstwa od tego tła uznaje się za obiekty poruszające się[221]. W bardziej zaawansowanych przypadkach wyodrębnia się punkty charakterystyczne w każdej klatce obrazu, korzystając z takich deskryptorów jak SIFT (ang. *Scale-Invariant*

[220] Zob. S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan and S. Kasaei, Deep Learning for Visual Tracking: A Comprehensive Survey, IEEE Transactions on Intelligent Transportation Systems, T. 23, nr 5, 2022, s. 3943-3968.

[221] Zob. M. Piccardi, Background subtraction techniques: a review, 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, Netherlands, 2004, s. 3099-3104.

Feature Transform)[222] lub SURF (ang. *Speeded-Up Robust Features*)[223]. Później wiąże się te punkty między kolejnymi klatkami, aby śledzić trajektorie obiektów[224]. Tego rodzaju rozwiązania bywają dobrze zrozumiałe i przejrzyste, można więc dość łatwo wskazać, dlaczego system „widział" w obrazie pewną krawędź bądź region. Z drugiej strony, metody te pozostają silnie zależne od dobrego doboru cech i często nie radzą sobie z większą zmiennością oświetlenia, skali, obrotów czy zasłonięcia. W nowoczesnych metodach opartych na uczeniu głębokim kluczową rolę odgrywają wielowarstwowe sieci neuronowe[225]. Zamiast ręcznego projektowania zestawu cech istotnych dla zadania detekcji bądź klasyfikacji obiektów w obrazie, współczesne modele głębokiego uczenia opierają się na mechanizmach automatycznego pozyskiwania reprezentacji cech bezpośrednio z danych wejściowych. Proces ten realizowany jest w toku treningu na dużych zbiorach danych, w których model uczy się hierarchii reprezentacji – od niskopoziomowych cech (takich jak krawędzie, tekstury) po bardziej złożone i semantycznie znaczące struktury (np. kontury obiektów czy ich przestrzenno-czasowe konfiguracje). W analizie materiałów wideo powszechnie stosuje się konwolucyjne sieci neuronowe oraz ich rozszerzenia uwzględniające wymiar czasowy, takie jak trójwymiarowe sieci konwolucyjne 3D-CNN (ang. *Three-Dimensional Convolutional Neural Networks*) lub rekurencyjne architektury LSTM (ang. *Long Short-Term Memory*) i GRU (ang. *Gated Recurrent Unit*), które umożliwiają analizę sekwencyjną klatek wideo[226,227]. Podejścia te, określane często mianem *end-to-end*, realizują bezpośrednie mapowanie sygnału wejściowego (ciągu klatek wideo) na wyjściową decyzję klasyfikacyjną lub detekcyjną, bez konieczności implementacji pośrednich etapów, takich jak

[222] Zob. D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60(2), 91-110.

[223] Zob. H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded Up Robust Features, Computer Vision and Image Understanding, 110(3), 404-417.

[224] Zob. A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys, 38(4), 13-45.

[225] Zob. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature, 521(7553), 436-444.

[226] Zob. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015, s. 2625-2634.

[227] Zob. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, s. 4489-4497.

ręczna ekstrakcja cech czy konstrukcja dedykowanych heurystyk. Z punktu widzenia instytucji odpowiedzialnych za bezpieczeństwo publiczne i narodowe, takie podejście umożliwia skuteczną identyfikację zdarzeń niebezpiecznych oraz automatyzację procesów obserwacyjnych w środowiskach o dużej dynamice i złożoności. Należy jednak zaznaczyć, że efektywność modeli *end-to-end* jest silnie uzależniona od dostępności odpowiednio dużych, reprezentatywnych zbiorów treningowych oraz od wysokiej wydajności zasobów obliczeniowych, co w warunkach operacyjnych może stanowić istotne ograniczenie. W tabeli 4 przedstawiono zestawienie kluczowych różnic pomiędzy klasycznymi a nowoczesnymi metodami analizy danych wizyjnych, z uwzględnieniem ich implikacji dla zastosowań w kontekście systemów bezpieczeństwa.

Tabela 4. Klasyczne a nowoczesne metody analizy wideo – kluczowe różnice

Aspekt	Klasyczne metody	Nowoczesne metody z zastosowaniem *deep learning*
Źródło cech	Ręcznie definiowane (np. krawędzie, punkty kluczowe, histogramy gradientów).	Automatyczne uczenie się cech z danych (sieci CNN, RNN, 3D-CNN).
Architektura	Modułowa (detekcja ruchu, segmentacja, ekstrakcja cech, klasyfikacja).	*End-to-end* – jedna spójna sieć ucząca się całego zadania (np. wykrycie + klasyfikacja w jednym).
Wymagania co do danych	Stosunkowo niewielkie zbiory treningowe.	Bardzo duże bazy danych, konieczność różnorodnego materiału (np. z wieloma wariantami tła).
Złożoność obliczeniowa	Zwykle niższa, możliwa implementacja nawet w urządzeniach o małej mocy.	Wysoka, wymaga procesorów GPU/TPU do trenowania, optymalizacje do pracy w czasie rzeczywistym.
Dokładność	Zależy silnie od jakości ręcznie zaprojektowanych cech, często niższa przy dużej zmienności.	Zazwyczaj wyższa i lepiej uogólniająca, zwłaszcza w zadaniach złożonych, o ile dostępne są bogate dane treningowe.
Odporność na zakłócenia	Metody dedykowane wąskim scenariuszom (możliwa utrata wydajności przy zmianie warunków).	Możliwość trenowania na rozmaitych warunkach (dzień/noc, różne kąty widzenia, tłok), co przekłada się na większą adaptacyjność.
Wyjaśnialność	Bardziej zrozumiałe (łatwe do interpretacji, dlaczego dana cecha była wykryta).	Trudniejsze do wyjaśnienia (model głębokiej sieci jest tzw. czarną skrzynką), choć rozwijane są metody XAI (ang. *explainable* AI).
Adaptacja do nowych zadań	Konieczność przeprojektowania cech i heurystyk.	Dzięki transfer learning możliwe szybkie dopasowanie do nowej domeny (np. *fine-tuning* sieci).
Przykładowe algorytmy	*Background subtraction*, *Kanade–Lucas–Tomasi* (KLT), SURF/SIFT + klasyfikator SVM.	YOLO (ang. *You Only Look Once*), *Faster* R-CNN, 3D-*ResNet*, LSTM/GRU do analizy sekwencji czasowych.

Źródło: opracowanie własne.

Początkowe zastosowania klasycznych metod analizy wideo, realizowane w warunkach ograniczonych zasobów obliczeniowych, przyniosły wymierne rezultaty szczególnie w zadaniach o ograniczonej skali, takich jak monitorowanie niewielkich obszarów czy detekcja ruchu w statycznym tle[228]. Systemy te konstruowano w oparciu o modularną architekturę, obejmującą kolejne etapy: wykrywanie obszarów ruchu, segmentację, ekstrakcję cech i śledzenie. Struktura taka zapewniała wysoką interpretowalność działania każdego z modułów. Niemniej jednak, błędy powstające na jednym etapie systemu miały tendencję do propagowania się, co skutkowało obniżeniem ogólnej efektywności rozpoznania. W szczególności, metody klasyczne wykazywały znaczną wrażliwość na zmienność warunków środowiskowych – dynamiczne oświetlenie (np. przejściowe zacienienia, odbicia światła), zmiany perspektywy kamery czy złożoność sceny związanej z obecnością tłumu znacząco obniżały skuteczność detekcji i klasyfikacji[229]. Rozwój technologii głębokiego uczenia zapoczątkował przejście od modularnych struktur analitycznych do ujednoliconych architektur typu *end-to-end*. W tego rodzaju podejściu sieć neuronowa przetwarza bezpośrednio surowy strumień danych wizyjnych, optymalizując swoje wewnętrzne reprezentacje cech w celu maksymalizacji skuteczności detekcji i klasyfikacji obiektów. Wstępne warstwy konwolucyjne uczą się wychwytywać niskopoziomowe cechy strukturalne, takie jak kontury czy krawędzie, natomiast kolejne warstwy agregują i przekształcają te informacje w bardziej abstrakcyjne reprezentacje semantyczne[230,231]. Wyższe warstwy sieci mogą wykrywać określone klasy obiektów (np. sylwetki ludzi, pojazdy) oraz złożone wzorce ruchowe, pozwalające na rozróżnienie aktywności, takich jak marsz, bieg czy pozostawienie przedmiotu. W licznych benchmarkach badawczych nowoczesne metody oparte na architekturach głębokiego uczenia przewyższają skutecznością klasyczne

[228] Zob. T. Zhang, Z. Liu, X. Lian, X. Wang, Study on moving-objects detection technique in video surveillance system, Chinese Control and Decision Conference, Xuzhou, 2010, s. 2375-2380.

[229] Zob. A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys, 38(4), s. 13-45.

[230] Zob. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in Neural Information Processing Systems, 27, 2014, s. 568-576.

[231] Zob. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, s. 4489-4497.

rozwiązania. Na przykład w zbiorze MS COCO[232], zawierającym obiekty w zróżnicowanych kontekstach przestrzennych i semantycznych, modele klasyczne osiągały wskaźnik mAP (ang. *mean Average Precision*) na poziomie kilkudziesięciu procent, podczas gdy współczesne modele głębokie (np. YOLOv3 czy Mask R-CNN) przekraczają 70–80% mAP[233,234]. W kontekście zastosowań bezpieczeństwa takie rezultaty oznaczają znacznie większą trafność w wykrywaniu nietypowych zdarzeń, identyfikacji osób poszukiwanych czy klasyfikacji podejrzanych aktywności w przestrzeni publicznej. Równolegle istotnym czynnikiem pozostaje szybkość działania analizatorów, zwłaszcza w kontekście ograniczeń zasobów obliczeniowych. Klasyczne metody, w tym detektory ruchu oparte na różnicach w kolejnych klatkach (np. *background subtraction*), mogą być implementowane bezpośrednio na urządzeniach brzegowych, takich jak kamery z ograniczonym procesorem[235]. W takich systemach możliwe jest realizowanie filtracji wstępnej na poziomie urządzenia, a przesyłanie danych do bardziej złożonej analizy następuje jedynie po spełnieniu określonych warunków. Z kolei modele głębokiego uczenia wymagają zwykle stosowania dedykowanych akceleratorów, takich jak procesory graficzne lub jednostki obliczeniowe TPU (ang. *Tensor Processing Unit)*. W odpowiedzi na te ograniczenia, rozwijane są modele zoptymalizowane pod kątem wydajności i rozmiaru, np. *MobileNet* czy *SqueezeNet,* umożliwiające stosowanie analizy głębokiej w warunkach brzegowych[236,237]. Odporność modeli na zmienność warunków środowiskowych stanowi kolejny ważny aspekt porównawczy. Algorytmy klasyczne często projektowano z myślą

[232] Zob. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft COCO: Common Objects in Context, arXiv preprint arXiv:1405.0312v3, 2015.

[233] Zob. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement, arXiv preprint arXiv:1804.02767, 2018.

[234] Zob. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, 2017, s. 2961-2969.

[235] Zob. M. Piccardi, Background subtraction techniques: a review, 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), The Hague, Netherlands, 2004, s. 3099-3104.

[236] Zob. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, arXiv preprint arXiv:1704.04861, 2017.

[237] Zob. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, arXiv preprint arXiv:1602.07360, 2016.

o statycznych scenariuszach, takich jak detekcja ruchu na tle o stałym kontraście, co czyniło je mało elastycznymi w konfrontacji z realnymi zmianami warunków oświetleniowych, pogodowych lub dynamicznych ujęć. Architektury głębokiego uczenia, trenowane na zróżnicowanych zbiorach danych (obejmujących m.in. sceny dzienne, nocne, zatłoczone lub zamglone), wykazują wyraźnie większy stopień generalizacji. Warto podkreślić możliwość wykorzystania metod transferu wiedzy (ang. *transfer learning*), polegających na adaptacji wcześniej wytrenowanych modeli do nowego środowiska operacyjnego poprzez *fine-tuning*. Pozwala to na szybkie wdrożenie detektorów przystosowanych do specyficznych warunków, np. w portach lotniczych, na przejściach granicznych czy w systemach ochrony infrastruktury krytycznej. Niewątpliwym ograniczeniem architektur typu *end-to-end* pozostaje ich niska interpretowalność. W tradycyjnych systemach wieloetapowych możliwe jest dokładne wskazanie, który komponent odpowiada za określony etap przetwarzania, a w przypadku błędów – szybkie zlokalizowanie ich źródła. Tymczasem w głębokich modelach neuronowych granica między kolejnymi warstwami często ulega zatarciu, a sam proces decyzyjny jest nieprzejrzysty. Trudności te są szczególnie istotne w kontekście analizy dowodowej, gdzie konieczne może być uzasadnienie działania algorytmu przed sądem lub organem nadzorczym[238]. Z tego względu rozwijane są narzędzia XAI (ang. *Explainable Artificial Intelligence*), których zadaniem jest wizualizacja aktywności sieci oraz identyfikacja obszarów klatki wideo odpowiedzialnych za wywołanie decyzji klasyfikacyjnej[239]. Praktyczną ilustracją różnic między podejściami może być scenariusz monitorowania ruchu drogowego na skrzyżowaniu. W podejściu klasycznym zastosowanie znalazłyby sensory indukcyjne lub czujniki ruchu do wykrywania pojazdów, algorytmy HOG (ang. *Histograms of Oriented Gradient*) do detekcji sylwetek pieszych[240] oraz filtr Kalmana do śledzenia obiektów w kolejnych klatkach[241]. System taki, choć

[238] Zob. S. Gu, L. Rigazio, Towards deep neural network architectures robust to adversarial examples, ICLR 2015 Workshop, 2015.

[239] Zob. W. Samek, T. Wiegand, K.R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, ITU Journal: ICT Discoveries, 1(1), 2017, s. 39-48.

[240] Zob. N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, 2005, s. 886-893.

[241] Zob. R.E. Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering, 82(1), 1960, s. 35-45.

funkcjonalny, wykazywałby wrażliwość na zasłonięcia obiektów lub zakłócenia świetlne. Tymczasem nowoczesne sieci głębokiego uczenia, jak YOLOv5, są zdolne do jednoczesnej detekcji wielu klas obiektów, ich identyfikacji oraz śledzenia trajektorii – również w warunkach częściowego przesłonięcia czy zróżnicowanego oświetlenia.

W zastosowaniach instytucji bezpieczeństwa państwa obie grupy metod – klasyczne i głębokie – są nadal wykorzystywane w sposób komplementarny. W wielu przypadkach wdrażana jest architektura hybrydowa, w której algorytmy klasyczne realizują wstępną analizę lub filtrację, a modele głębokie stosowane są jedynie w odniesieniu do wykrytych fragmentów o podwyższonym ryzyku. Pozwala to ograniczyć wymagania sprzętowe, zredukować obciążenie transmisyjne oraz skrócić czas reakcji w sytuacjach incydentalnych. Technologiczny rozwój w ostatnich latach doprowadził również do pojawienia się bardziej złożonych modeli 3D-CNN, uwzględniających informację temporalną w rozpoznawaniu aktywności[242], co otworzyło nowe możliwości w zakresie detekcji zdarzeń dynamicznych, takich jak gwałtowny bieg, upadek czy bójka. Zagadnieniem budzącym rosnące zainteresowanie jest również bezpieczeństwo samych modeli SI w kontekście tzw. ataków przeciwnych (ang. *adversarial attacks*). Niewielkie, celowe modyfikacje pikseli obrazu – niedostrzegalne dla ludzkiego oka – mogą doprowadzić do całkowitej zmiany decyzji modelu, np. sklasyfikowania uzbrojonego napastnika jako neutralnego obiektu[243]. Z tego względu wdrażane są strategie zabezpieczające modele przed tego rodzaju manipulacjami, a także hybrydowe architektury oparte na współdziałaniu klasyfikatorów głębokich i reguł eksperckich. W kontekście systemów krytycznych (ang. *mission-critical*), takich jak nadzór nad infrastrukturą strategiczną, istotne staje się formułowanie dodatkowych reguł decyzyjnych, które zwiększają interpretowalność i redukują liczbę fałszywych alarmów[244]

Podsumowując, klasyczne metody analizy wideo oferują prostotę, niskie wymagania obliczeniowe i wysoką przejrzystość działania, przez co

[242] Zob. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, s. 4489-4497.

[243] Zob. A. Kurakin, I. Goodfellow, S. Bengio, Adversarial machine learning at scale, 5th International Conference on Learning Representations (ICLR), Toulon, 2017, s. 1-17.

[244] Zob. Y. Wang, S.H. Chung, Artificial intelligence in safety-critical systems: a systematic review, Industrial Management & Data Systems, 121(9), 1999-2024.

pozostają przydatne w określonych zastosowaniach o stabilnym środowisku i ograniczonej skali. Natomiast metody głębokiego uczenia zapewniają znacznie wyższą skuteczność, adaptacyjność do złożonych warunków operacyjnych i możliwość zastosowania w architekturze *end-to-end*, przy jednoczesnej potrzebie intensywnych zasobów danych i obliczeń. Wraz z rozwojem infrastruktury obliczeniowej, dostępnością narzędzi takich jak GPU i TPU, a także ekspansją platform chmurowych i rozwiązań *edge computing*, integracja analizatorów głębokiego uczenia z systemami bezpieczeństwa nabiera coraz większego znaczenia. Z perspektywy nauk o bezpieczeństwie wybór metody powinien uwzględniać nie tylko skuteczność, ale także przewidywalność, koszty wdrożenia, czas reakcji oraz możliwość transparentnego wyjaśnienia decyzji systemu.

6.3. Integracja systemów wykrywania dezinformacji z instytucjami bezpieczeństwa

W skali globalnej rozwinięto szereg rozwiązań integrujących systemy wykrywania dezinformacji z instytucjami odpowiedzialnymi za bezpieczeństwo państwa. W Stanach Zjednoczonych do czołowych inicjatyw należało utworzenie wyspecjalizowanych jednostek analitycznych i koordynacyjnych. Przykładem jest *Global Engagement Center* (GEC) w Departamencie Stanu USA, które pełniło rolę centrum nerwowego amerykańskiej strategii przeciwdziałania dezinformacji ze strony obcych państw. Zadaniem GEC (utworzonego w 2016 r. i rozszerzonego ustawowo w 2017 r.) była koordynacja międzyagencyjna w ujawnianiu i przeciwdziałaniu propagandzie oraz kampaniom informacyjnym prowadzonym przez wrogie podmioty. Centrum to realizowało działania w pięciu obszarach: prowadziło analizy danych o operacjach informacyjnych przeciwnika, rozwijało partnerstwa międzynarodowe, inicjowało programy tematyczne (np. przeciw dezinformacji z Rosji, Chin, Iranu), koordynowało ujawnianie wpływowych operacji informacyjnych oraz testowało nowe technologie do wykrywania fałszywych treści. Dodatkowo, amerykańska społeczność wywiadowcza w 2022 r. powołała *Foreign Malign Influence Center* (FMIC) przy Dyrektorze Wywiadu Narodowego, integrujące dane wywiadowcze o zagrożeniach dezinformacyjnych i wspierające takie jednostki jak GEC. Równolegle Departament Bezpieczeństwa Wewnętrznego

(DHS) oraz agencje wywiadowcze i wojskowe rozwijają własne komórki analizy informacji (np. w ramach *Cybersecurity and Infrastructure Security Agency* do ochrony wyborów czy w ramach dowództw zajmujących się operacjami informacyjnymi). Mimo rozproszenia, Stany Zjednoczone dążyły do integracji tych wysiłków – raport GAO z 2024 r. wskazywał na konieczność usprawnienia współpracy i utworzenia spójnych standardów w całym rządzie. W Unii Europejskiej od 2015 r. znacznie rozwinęły się zdolności instytucjonalne do przeciwdziałania dezinformacji. Na szczeblu unijnym kluczową rolę odgrywa *European External Action Service* (EEAS) i jego specjalny zespół *East StratCom Task Force*, znany z platformy *EUvsDisinfo*. Jego zadaniem jest stałe monitorowanie i ujawnianie przypadków dezinformacji (zwłaszcza pochodzącej z Rosji) oraz zwiększanie świadomości publicznej na ten temat. UE wprowadziła również Kodeks postępowania w zakresie dezinformacji – dobrowolne porozumienie z platformami internetowymi (od 2018, zaktualizowane w 2022), które stało się prekursorem regulacji zobowiązujących media społecznościowe do współpracy w identyfikowaniu i ograniczaniu rozprzestrzeniania fałszywych treści. Najnowsze regulacje, takie jak Akt o usługach cyfrowych (DSA) z 2022 r., nakładają na bardzo duże platformy obowiązek oceny i ograniczania systemowych zagrożeń dla swoich użytkowników – w tym ryzyka związanego z dezinformacją. Na poziomie państw członkowskich podejmowano inicjatywy prawne: np. Francja przyjęła w 2018 r. prawo przeciw manipulacjom informacyjnym w okresie wyborczym, a Niemcy rozszerzyły przepisy (NetzDG) zobowiązujące platformy do szybkiego usuwania nie tylko mowy nienawiści, ale i szkodliwych *fake newsów*. Ponadto w 2019 r. zainicjowano unijny System Wczesnego Ostrzegania przed dezinformacją, aby usprawnić wymianę informacji między państwami UE. Działania UE są ściśle powiązane z współpracą międzynarodową – UE kooperuje m.in. z NATO, G7 (*Rapid Response Mechanism*), ONZ (kampania *Verified*) oraz centrami analitycznymi USA, tworząc szeroki front przeciw manipulacjom informacyjnym. NATO traktuje dezinformację jako element zagrożeń hybrydowych dla bezpieczeństwa Sojuszu. W strukturach NATO rozwijane są zdolności tzw. Strategicznej Komunikacji (*StratCom*), integrujące działania z zakresu informacji i komunikacji na poziomie wojskowym i politycznym. W Rydze działa Centrum Eksperckie NATO ds. Komunikacji Strategicznej (*NATO StratCom COE*), prowadzące badania nad metodami wykrywania i przeciwdziałania propagandzie oraz szkolące personel państw

członkowskich w tym zakresie. NATO blisko współpracuje z UE (np. poprzez wymianę analiz z EEAS) oraz krajami partnerskimi w zakresie rozpoznawania kampanii dezinformacyjnych. W 2024 r. NATO zaktualizowało swoją strategię AI, uwzględniając zagrożenia związane z dezinformacją wspieraną przez sztuczną inteligencję. Pojawiają się inicjatywy tworzenia wspólnego *hubu* wywiadu informacji – proponowany NATO *Information Threat Intelligence Hub* miałby z wykorzystaniem AI w czasie rzeczywistym wykrywać złośliwe sieci szerzące dezinformację i umożliwiać ich szybkie neutralizowanie. Jednocześnie NATO prowadzi ćwiczenia obrony przed złożonymi atakami, łączącymi cyberoperacje z kampaniami informacyjnymi, aby wypracować procedury reagowania na tzw. ataki poniżej progu artykułu 5. Sojuszu . Integracja systemów wykrywania dezinformacji następuje także poprzez tworzenie wielonarodowych zespołów – np. w 2023 r. zainicjowano koncepcję *Global Counter-Hybrid Alliance* (NATO, UE i m.in. Australia, Japonia) do wymiany ostrzeżeń i skoordynowanych reakcji na kampanie informacyjne . W innych państwach również powstają wyspecjalizowane struktury łączące kompetencje analityczne z bezpieczeństwem. Wielka Brytania utworzyła w 2019 r. rządową jednostkę *Counter Disinformation Unit* (CDU) w celu identyfikowania i przeciwdziałania kampaniom wpływu, zwłaszcza ze strony obcych państw. Jednostka ta (obecnie funkcjonująca jako *National Security and Online Information Team* przy Departamencie ds. Nauki, Innowacji i Technologii) monitoruje otwarte źródła, śledzi narracje i trendy w sieci oraz próby sztucznej manipulacji środowiskiem informacyjnym. Brytyjskie CDU współpracuje z wywiadem i organami ścigania, a także z firmami z sektora AI – rząd zlecał np. firmie Logically analizy dezinformacji online. Istnienie takiej jednostki wywołuje jednak dyskusje etyczne – konieczne jest zachowanie równowagi między bezpieczeństwem informacyjnym a wolnością słowa. Finlandia natomiast przyjęła podejście systemowe oparte na budowaniu odporności społecznej: jest powszechnie uznawana za modelowe państwo w zwalczaniu dezinformacji dzięki połączeniu bardzo wysokiej wolności prasy z nowatorską edukacją medialną społeczeństwa. Fińska strategia od lat stawia na edukację od najmłodszych lat i promocję krytycznego myślenia – Finlandia osiąga najwyższe wskaźniki kompetencji medialnych w Europie, co przekłada się na odporność na wpływy informacyjne. Jednocześnie władze fińskie monitorują zagrożenia (szczególnie z Rosji), choć do niedawna formalne reakcje państwa były ograniczone – opierano się na silnym ekosystemie

niezależnych mediów i *fact-checkingu* (np. *NGO Faktabaari*) oraz zaufaniu społecznym. Przykłady te pokazują zróżnicowanie podejść: od scentralizowanych centrów analityczno-wywiadowczych (USA), przez struktury komunikacyjne NATO i UE, po działania regulacyjne i edukacyjne wzmacniające świadomość społeczną (UE, Finlandia). Wspólnym trendem jest jednak coraz większa integracja technologii, struktur organizacyjnych, regulacji prawnych i działań edukacyjnych w celu stworzenia spójnego systemu bezpieczeństwa informacyjnego państwa.

Systemy detekcji dezinformacji mogą być integrowane z systemem bezpieczeństwa państwa na różnych poziomach. Należą do nich: narzędzia analityczne (technologie), struktury organizacyjne, ramy prawne i regulacyjne oraz działania edukacyjne i świadomościowe. Każdy z tych poziomów odgrywa odmienną rolę w całościowym ekosystemie bezpieczeństwa informacyjnego. Poniżej w tabeli 5 przedstawiono charakterystykę tych poziomów integracji wraz z przykładami działań.

Tabela 5. Charakterystyka poziomów integracji

Poziom integracji	Charakterystyka	Przykłady
Narzędzia analityczne (techniczny)	Wdrożenie zaawansowanych systemów informatycznych do automatycznego wykrywania, analizy i filtrowania treści fałszywych. Integracja algorytmów SI w infrastrukturę instytucji bezpieczeństwa.	Systemy AI skanujące media społecznościowe pod kątem *botów* i *fake news*; detektory *deepfake'ów* w agencjach wywiadowczych; panele analityczne w centrach operacyjnych służb.
Struktury organizacyjne (instytucjonalny)	Powołanie specjalnych komórek/zespołów w ramach istniejących instytucji bezpieczeństwa lub utworzenie nowych jednostek dedykowanych walce z dezinformacją. Zapewnienie kanałów współpracy między agencjami.	Jednostki typu GEC (USA) czy CDU/NSOIT (Wielka Brytania); zespoły analizy informacji w wywiadzie i armii; międzyresortowe zespoły reagowania kryzysowego ds. dezinformacji.
Ramy prawne i regulacyjne	Ustanowienie przepisów prawa i procedur umożliwiających instytucjom bezpieczeństwa monitorowanie i przeciwdziałanie dezinformacji z poszanowaniem praw obywatelskich. Tworzenie standardów i wytycznych etycznych dla działań w infosferze.	Ustawy wyborcze przeciwdziałające „*fake news*" (np. Francja 2018); regulacje UE jak DSA zobowiązujące platformy do ograniczania dezinformacji; wytyczne dot. wykorzystania SI zgodnie z prawami człowieka (projekt Traktatu o SI Rady Europy).
Edukacja i świadomość (społeczny)	Integracja zagadnień związanych z dezinformacją w systemie edukacji oraz szkoleniach służb. Budowanie odporności społecznej poprzez uświadamianie obywateli i urzędników o mechanizmach propagandy.	Edukcja medialna w szkołach (np. Finlandia od przedszkola po uniwersytet; kampanie informacyjne dla społeczeństwa (ONZ „*Verified*" 2020); szkolenia dla żołnierzy i funkcjonariuszy z rozpoznawania manipulacji.

Źródło: opracowanie własne.

Jak wynika z powyższego zestawienia, skuteczna integracja systemów wykrywających dezinformację wymaga synergii działań na wszystkich poziomach. Narzędzia analityczne dostarczają mocy technologicznej – dzięki nim możliwe jest przesiewanie ogromnych wolumenów danych i identyfikacja podejrzanych wzorców treści. Struktury organizacyjne zapewniają zaś formalne osadzenie tych zdolności w instytucjach bezpieczeństwa (poprzez dedykowane zespoły i procedury), co umożliwia praktyczne wykorzystanie wyników analiz w procesie decyzyjnym państwa. Ramy prawne tworzą podstawy do legalnego działania – regulują uprawnienia służb w zakresie monitoringu informacji, nakładają obowiązki na podmioty prywatne (np. operatorów platform internetowych) oraz ustanawiają mechanizmy kontroli i odpowiedzialności, aby zapobiec nadużyciom. Z kolei edukacja i budowanie świadomości to poziom fundamentalny i długofalowy – nawet najlepsze algorytmy nie zastąpią świadomego, krytycznie myślącego społeczeństwa i kompetentnych kadr. Dlatego też państwa inwestują w programy podnoszenia kompetencji medialnych obywateli, co w efekcie integruje społeczeństwo obywatelskie w szeroko rozumiany system bezpieczeństwa informacyjnego. Podział na powyższe poziomy ma charakter analityczny – w praktyce one się przenikają. Przykładowo, wdrożenie nowego narzędzia AI do wykrywania fałszywych treści wymaga zarówno utworzenia odpowiedniej komórki organizacyjnej, która będzie je obsługiwać, jak i uregulowania podstaw prawnych dla przetwarzania danych, a także przeszkolenia analityków i oficerów.

Integracja systemów wykrywania dezinformacji z instytucjami bezpieczeństwa państwa wymaga uwzględnienia specyfiki działania różnych służb: wywiadowczych, cyberbezpieczeństwa oraz wojskowych. Każda z tych instytucji wnosi odmienny kontekst operacyjny, co wpływa na wymagania techniczne i rozwiązania organizacyjne. Służby wywiadowcze i kontrwywiadowcze tradycyjnie zajmują się pozyskiwaniem i analizą informacji, dlatego naturalnym miejscem integracji systemów detekcji dezinformacji są struktury wywiadu cywilnego i wojskowego. W ostatnich latach wywiady rozwijają zdolności OSINT (ang. *Open Source Intelligence*) właśnie pod kątem monitorowania otwartych źródeł informacji, w tym mediów społecznościowych, forów internetowych czy serwisów informacyjnych. Systemy wykrywania dezinformacji dostarczają wywiadowi zautomatyzowanych narzędzi do filtrowania szumu informacyjnego i wychwytywania sygnałów wskazujących na skoordynowane kampanie wpływu. Technicznie oznacza

to integrację algorytmów analizujących treści tekstowe np. wykrywanie narracji propagandowych i obrazy czy wideo np. detekcja *deepfake* z istniejącymi platformami analitycznymi wywiadu. Organizacyjnie wiele państw utworzyło w strukturach wywiadowczych wyspecjalizowane komórki analizy informacji. Przykładowo, w ramach ODNI (ang. *Office of the Director of National Intelligence*) w USA działa *Foreign Malign Influence Center*, integrujące dane wielu agencji i dostarczające analitykom wywiadu skonsolidowany obraz zagrożeń z zakresu dezinformacji. Podobne funkcje pełnią wydziały w agencjach bezpieczeństwa odpowiadające za kontrpropagandę czy wojnę informacyjną. Integracja z wywiadem umożliwia, by dane z systemów detekcji (np. wykryte sieci kont botów) były od razu łączone z innymi informacjami wywiadowczymi (np. danymi o powiązaniach osób, strukturom finansowania itp.), co znacznie zwiększa wartość analityczną – pozwala wskazać, kto stoi za kampanią i jakie są jej cele. Wyzwania techniczne obejmują tutaj zapewnienie bezpieczeństwa i poufności danych, gdzie systemy muszą działać w wydzielonych, zabezpieczonych środowiskach IT oraz interoperacyjności, w ramach których narzędzia AI powinny móc korzystać zarówno z danych *open-source*, jak i niejawnych baz danych, przy zachowaniu odpowiednich ograniczeń dostępu. Agencje ds. cyberbezpieczeństwa np. rządowe centra reagowania na incydenty komputerowe, takie jak CERT/CSIRT coraz częściej włączają komponent informacyjny do swojej działalności. Dzieje się tak, ponieważ ataki cybernetyczne bywają sprzężone z kampaniami dezinformacyjnymi tzw. ataki hybrydowe. Przykładem jest równoczesne włamanie do infrastruktury i rozprzestrzenienie fałszywej narracji mającej wywołać panikę. Instytucje cyberbezpieczeństwa integrują więc systemy wykrywania dezinformacji, aby w ramach monitoringu zagrożeń sieciowych wychwytywać także groźne manipulacje informacyjne. Technicznie oznacza to rozszerzenie istniejących platform SIEM (ang. *Security Information and Event Management*) o moduły analizujące treści w mediach (np. wzmianki o awariach, atakach, katastrofach mogące być *fake news*. Organizacyjnie często tworzy się hybrydowe zespoły analityczne łączące ekspertów od cyberbezpieczeństwa z analitykami informacji. Przykładowo, przy NATO działają grupy robocze łączące specjalistów cyberobrony z ekspertami od komunikacji strategicznej, aby razem ćwiczyć scenariusze obrony przed równoczesnym atakiem cybernetycznym i dezinformacyjnym. Taka integracja wymaga wypracowania wspólnego języka

i procedur – analitycy cyber muszą rozumieć modele rozpowszechniania dezinformacji, a analitycy informacji – podstawy technik hakerskich, by móc skoordynować działania. Pod względem technicznym istotne jest zapewnienie szybkiej wymiany danych między systemami monitorowania sieci a systemami monitorowania mediów (np. alert o pojawieniu się dezinformacji dotyczącej ataku na infrastrukturę krytyczną powinien natychmiast trafić do sztabu zarządzania incydentem. Wymaga to integracji kanałów komunikacji i wspólnych platform raportowania. Siły zbrojne integrują zagadnienie dezinformacji zarówno na poziomie strategicznym, operacyjnym, jak i taktycznym. Na poziomie strategicznym wojska wielu państw (np. USA, Wielka Brytania, Polska) posiadają doktryny prowadzenia operacji informacyjnych i walki psychologicznej, które uwzględniają przeciwdziałanie propagandzie przeciwnika oraz prowadzenie własnych działań informacyjnych. Systemy wykrywania dezinformacji są tu wykorzystywane do sytuacyjnej świadomości informacyjnej – pozwalają dowództwu ocenić, jakie narracje są szerzone w rejonie konfliktu i jak mogą wpływać na morale wojsk własnych czy nastawienie ludności cywilnej. Dlatego np. w strukturach dowództw pojawiają się komórki analizy medialnej na teatrze działań. Technicznie wojsko wdraża narzędzia do monitorowania eteru informacyjnego w czasie zbliżonym do rzeczywistego – analizujące lokalne media, wpisy na platformach (Twitter/X, Telegram itp.) pod kątem dezinformacji związanej z operacjami militarnymi. Przykładem mogą być systemy używane podczas misji zagranicznych do śledzenia rozpowszechnianych plotek (np. fałszywe informacje o działaniach kontyngentu NATO) i ich natychmiastowego *debunkingu* we współpracy z sekcjami *public affairs.* W NATO wypracowano koncepcję *Resilient Narratives Task Force*, tj. zespołu szybkiego reagowania złożonego z ekspertów komunikacyjnych, analityków i sojuszniczych specjalistów, którego zadaniem byłoby w warunkach kryzysu szybko dostarczać wiarygodne kontr-narracje i prostować fałszywe informacje na polu walki informacyjnej. Integracja systemów detekcji z wojskiem wymaga także rozwiązań organizacyjnych na styku wojsko-cywilny: np. stałej współpracy z agencjami wywiadu cywilnego i mediami. Wojsko coraz częściej uczestniczy w narodowych strukturach reagowania na kryzysy informacyjne, a oficerowie są delegowani do międzyresortowych zespołów analizujących dezinformację w sytuacjach zagrożenia bezpieczeństwa państwa. Technicznie kluczowe jest dostosowanie narzędzi do wymagań pola walki – systemy

muszą działać w środowisku o ograniczonej łączności, być odporne na zakłócenia i bezpieczne przed przeciwnikiem by np. nie dały się zhakować i nie przekazywały fałszywych alarmów. Podsumowując, integracja z instytucjami bezpieczeństwa wymaga interdyscyplinarnego podejścia: połączenia nowoczesnych technologii analizy danych z odpowiednimi strukturami organizacyjnymi i procedurami działania. Każdy sektor – wywiad, cyber, wojsko – wnosi unikalną perspektywę, ale wszystkie korzystają z wspólnych zdobyczy technologicznych.

Sztuczna inteligencja odgrywa kluczową rolę w nowoczesnych systemach wykrywania dezinformacji. Ze względu na ogromną skalę danych generowanych każdego dnia w Internecie, ręczna analiza jest niewystarczająca – dlatego sięga się po algorytmy uczenia maszynowego do automatycznego wykrywania szkodliwych lub fałszywych treści. Wykorzystywane są różnorodne techniki sztucznej inteligencji, obejmujące m.in. przetwarzanie języka naturalnego NLP (ang. *Natural Language Processing*), metody klasyfikacji tekstu, algorytmy grafowe i detekcję anomalii. W przypadku analizy tekstu, modele NLP potrafią rozpoznawać charakterystyczne cechy językowe dezinformacji. Stosuje się np. analizę sentymentu oraz rozpoznawanie encji, czyli nazw własnych do wychwytywania podejrzanych wzorców w tekście. Treści dezinformacyjne często charakteryzują się skrajnym emocjonalnym zabarwieniem i powtarzalnymi frazami – nowoczesne modele językowe uczone na dużych zbiorach oznaczonych przykładów potrafią z pewnym prawdopodobieństwem odróżnić artykuł rzetelny od propagandowego. Klasyfikatory oparte na uczeniu głębokim (np. sieci neuronowe typu transformator, w tym modele podobne do BERT czy GPT) są w stanie kategoryzować informacje jako prawdziwe lub fałszywe na podstawie wzorców wyuczonych z danych treningowych. Jednocześnie wykorzystuje się algorytmy grafowe do analizy sieci rozpowszechniania informacji – traktując profile, strony i wiadomości jako graf połączeń, można wykrywać skoordynowane kampanie np. klastery kont powiązane ze sobą, typowe dla operacji wpływu). Dodatkowo, algorytmy wykrywania anomalii flagują nietypowe wzorce aktywności, które mogą wskazywać na występowanie zorganizowanej akcji dezinformacyjnej – przykładem może być nagły przyrost nowych kont rozpowszechniających ten sam przekaz lub nietypowe natężenie określonej narracji w krótkim czasie. Poza klasycznymi mediami tekstowymi, rośnie znaczenie analizy treści audiowizualnych. Multimodalne systemy detekcji łączą analizę obrazu,

dźwięku i tekstu, by wykrywać fałszerstwa w materiałach wideo, audio czy grafikach. Rozwojowi uległy narzędzia do wykrywania *deepfake'ów* – czyli syntetycznie generowanych lub zmanipulowanych materiałów wideo i audio, które mogą służyć dezinformacji. Wykorzystuje się tu sieci neuronowe (np. konwolucyjne sieci do analizy obrazów klatka po klatce), które uczą się subtelnych artefaktów pozostawianych przez algorytmy generujące deepfake np. nieregularne mrugnięcia, zniekształcenia krawędzi obiektów, nienaturalne ruchy ust względem mowy. Ponadto bada się cechy biometryczne – np. rytm pulsacji naczyń krwionośnych na twarzy czy mikroekspresje – jako wskaźniki odróżniające prawdziwe nagranie od sztucznie wygenerowanego. W efekcie najlepsze detektory potrafią osiągać wysoką skuteczność w identyfikacji *deepfake*[245]. Trzeba jednak podkreślić, że trwa wyścig zbrojeń między twórcami *deepfake* a twórcami detektorów – wraz z postępem generatywnej AI np. modeli typu GAN, fałszerstwa stają się trudniejsze do wychwycenia, co wymusza ciągłą aktualizację i doskonalenie algorytmów wykrywających. Prócz wykrywania samych fałszywych treści, narzędzia AI wspomagają analizę metadanych i kontekstu. Wykorzystuje się np. analizę sieci społecznościowych do ustalania, jak dana informacja się rozprzestrzenia, kto jest źródłem pierwotnym, jakie boty lub fałszywe konta biorą udział w amplifikacji przekazu. Algorytmy potrafią grupować konta o podobnych wzorcach aktywności i wskazywać tzw. *coordinated inauthentic behawior*. Takie analizy grafowe są dziś standardowo integrowane z platformami analitycznymi instytucji bezpieczeństwa. Kolejnym aspektem jest automatyczna weryfikacja faktów – pojawiają się prototypy systemów, które na wejściu otrzymują artykuł lub wypowiedź i przeszukują wiarygodne bazy danych oraz źródła, by ocenić prawdziwość stwierdzeń. Choć pełna automatyczna *fact-checking* jest nadal wyzwaniem ze względu na kontekst i złożoność języka, to pewne elementy – jak sprawdzanie czy przypisane cytaty rzeczywiście padły, czy zdjęcia nie są użyte w kontekście oderwanym od oryginalnego (tzw. *context mismatch*) – mogą być w dużym stopniu zautomatyzowane. Jakość danych treningowych ma krytyczne znaczenie dla skuteczności tych modeli. Jeśli modele AI uczone są na stronniczych lub ograniczonych danych, mogą ulegać stronniczości i albo nie wykrywać

[245] Zob. B. Pinhasov, R. Lapid, R. Ohayon, M. Sipper, Y., Aperstein, Attacks on Deepfake Detectors, Afeka – The Academic College of Engineering in Tel Aviv & Ben-Gurion University of the Negev & DeepKeep, 2024.

określonych rodzajów dezinformacji, albo generować fałszywe alarmy dla treści prawdziwych, ale nietypowych. Dlatego prowadzi się inicjatywy kuratorskie – np. tworzenie obszernych, zrównoważonych zbiorów danych oznaczonych pod kątem dezinformacji. Istnieją bazy wiedzy i *datasety* jak *Media Bias/Fact Check, FakeNewsNet* itp., które są wykorzystywane do trenowania i testowania algorytmów. Równolegle rozwijane są metody oceny modeli w różnych warunkach, np. poprzez testy sprawiedliwości algorytmów, czy nie dyskryminują jakiejś grupy poglądów oraz audyty algorytmiczne. Tego typu działania mają zapewnić, że zintegrowane systemy detekcji działają rzetelnie i wiarygodnie również w zmiennych kontekstach. Należy zaznaczyć, że algorytmy wykrywające dezinformację są wspierane także przez narzędzia skierowane do użytkowników końcowych. Przykładem są rozszerzenia do przeglądarek i aplikacje mobilne (np. *NewsGuard, Fakeslayer*), które ostrzegają internautów przed niepewnymi źródłami lub analizują w czasie rzeczywistym wiarygodność czytanego artykułu. Takie narzędzia korzystają z zaplecza AI np. oceny reputacji witryny, analizy treści i stanowią pomost między czysto technicznym systemem a poziomem edukacji użytkownika. Podsumowując, integracja AI z systemami bezpieczeństwa umożliwiła automatyzację i przyspieszenie wykrywania dezinformacji na niespotykaną dotąd skalę. Algorytmy AI, w tym głębokie sieci neuronowe, są dziś w stanie analizować język, obrazy i zależności sieciowe szybciej niż zespoły ludzkie, dając służbom wczesne ostrzeżenia o potencjalnych zagrożeniach. Jednakże skuteczność tych systemów zależy nie tylko od samych algorytmów, ale również od ich właściwego użytkowania i zaufania, co prowadzi do kwestii udziału człowieka i wiarygodności modeli AI.

Włączenie zaawansowanych algorytmów AI do procesów decyzyjnych instytucji bezpieczeństwa rodzi pytania o ich niezawodność, zrozumiałość i odporność na manipulacje. Dlatego współczesne podejścia podkreślają rolę czynnika ludzkiego w pętli (ang. *human-in-the-loop*), potrzebę wyjaśnialności AI (XAI) oraz zapewnienie bezpieczeństwa modeli przed atakami przeciwnika. Podejście *human-in-the-loop* zakłada ścisłą współpracę między automatycznymi systemami detekcji a analitykami-ludźmi. Żadne narzędzie nie powinno działać całkowicie autonomicznie w izolacji – kluczowe decyzje (np. oznaczenie czegoś jako dezinformacja i podjęcie działań) powinny być weryfikowane lub nadzorowane przez kompetentnych analityków. Model ten wynika z faktu, że algorytmy choć szybkie, mogą popełniać błędy – np. nie

rozumieć kontekstu kulturowego wypowiedzi lub ironii – co człowiek jest w stanie wychwycić. Badania wskazują, że integracja ekspertów (np. *fact-checkerów*) z systemami AI znacząco podnosi dokładność wykrywania fałszu i zmniejsza liczbę fałszywych alarmów[246]. Człowiek w pętli może pełnić różne role: od etapu trenowania modeli (dostarczając zweryfikowane dane i poprawiając błędne klasyfikacje), poprzez monitoring bieżących wyników (flagi algorytmu są oceniane przez analityka przed podjęciem interwencji), aż po ulepszanie systemu (*feedback* od analityka służy do dalszego dopasowania algorytmu). Takie iteracyjne sprzężenie zwrotne powoduje, że system „uczy się" na podstawie wskazań człowieka, a człowiek korzysta z podsuwanych przez system sugestii – tworząc efektywny mechanizm łączący szybkość maszyn z intuicją i wiedzą ekspercką ludzi. Wdrożenie *human-in-the-loop* wiąże się jednak z wyzwaniami organizacyjnymi: wymaga zapewnienia odpowiedniej liczby przeszkolonych analityków, klarownych procedur eskalacji (kiedy decyzja ma przejść z maszyny do człowieka) oraz narzędzi interakcji (przyjazne interfejsy, które prezentują wyniki AI w zrozumiały sposób). Kwestia wyjaśnialności sztucznej inteligencji XAI jest ściśle związana z powyższym. Wyjaśnialność oznacza zdolność systemu AI do przedstawienia przejrzystego uzasadnienia swoich decyzji lub rekomendacji. W kontekście bezpieczeństwa to szczególnie istotne – analityk czy decydent musi rozumieć, dlaczego dany artykuł został oflagowany jako potencjalnie fałszywy przez model. Metody XAI obejmują techniki takie jak LIME czy SHAP, które potrafią wskazać, które cechy wejściowe (słowa, frazy, cechy obrazu) najmocniej wpłynęły na decyzję modelu. Dzięki temu użytkownik może ocenić, czy system kierował się sensownymi przesłankami, czy np. artefaktami niezwiązanymi z istotą sprawy. Wyjaśnialność buduje zaufanie do systemu – organizacje wdrażające AI podkreślają, że rzucenie światła na „czarną skrzynkę" algorytmu zwiększa akceptację i zaangażowanie użytkowników. Jest to kluczowe, by np. analitycy wywiadu chcieli korzystać z podpowiedzi AI, zamiast je ignorować jako niezrozumiałe. Ponadto, XAI pomaga we wczesnym wykrywaniu błędów i uprzedzeń modelu – jeśli wyjaśnienia wskazują, że model zwraca uwagę na nieistotne lub niepokojące czynniki np. systematycznie oflagowuje treści jednej opcji politycznej, można odpowiednio zareagować. W praktyce, budowanie wyjaśnialności staje się standardem

[246] Zob. D. Akinyele, G. Olaoye, Human-in-the-Loop Misinformation Detection: Collaborative Approach with Human Fact-Checkers, ResearchGate, 2024.

– od poziomu projektowania modeli, gdzie wybiera się architektury bardziej interpretowalne, jeśli to możliwe po poziom polityk. W kontekście integracji z instytucjami bezpieczeństwa pojawiają się wręcz regulacje rekomendujące stosowanie XAI. Przykładowo, organy doradcze w USA i Wielkiej Brytanii zalecają zapewnienie transparentności algorytmów używanych do celów bezpieczeństwa narodowego. Trzecim kluczowym zagadnieniem jest bezpieczeństwo samych modeli AI. Paradoksalnie, systemy stworzone do wykrywania zagrożeń informacyjnych mogą stać się celem wyspecjalizowanych ataków ze strony przeciwnika. Mowa tu o tzw. atakach przeciwstawnych ang. (*adwersarialnych*) na modele uczenia maszynowego. Przeciwnik może próbować zmylić algorytm detekcji, tak by nie wykrył on dezinformacji. W przypadku klasyfikatorów tekstu istnieją metody polegające na drobnych modyfikacjach treści np. podmiana kilku liter, czy zużycie synonimów w kluczowych zdaniach, które dla człowieka nie zmieniają znaczenia przekazu, ale mogą drastycznie obniżyć skuteczność modelu – model „nie rozpozna" znanego wzorca, bo został on celowo zmodyfikowany. W przypadku obrazów czy wideo mogą to być niewidoczne dla oka ludzkiego zaburzenia pikseli np. szum dodany do każdej klatki, które powodują, że detektor *deepfake* zostaje oszukany. Badania potwierdzają, że modele wykrywania *fake newsów* czy *deepfake'ów* są podatne na takie ataki – np. można tak zmanipulować dane wejściowe, by detektor zaklasyfikował fałszywą informację jako prawdziwą[247]. Co więcej, pojawiła się koncepcja „odwrócenia wyjaśnialności" – paradoksalnie techniki XAI mogą zostać użyte przez atakującego do poznania kryteriów, jakimi kieruje się model, a następnie stworzenia spreparowanych treści, które celowo ominą te kryteria[248]. Z tego powodu kwestie odporności modeli są przedmiotem intensywnych badań. Opracowuje się metody obrony, np. trenowanie modeli na danych zawierających również przykłady ataków, by nauczyć je odporności (ang. *adversarial training*) czy systemy meta-detekcji sygnalizujące, że dany *input* mógł być spreparowany adwersarialnie. W kontekście integracji z instytucjami bezpieczeństwa, zabezpieczenie modeli ma wymiar praktyczny: przeciwnik dysponujący

[247] Zob. F. Siciliano, L. Maiano, L. Papa, F. Baccini, I. Amerini, F. Silvestri, Adversarial Data Poisoning for Fake News Detection: How to Make a Model Misclassify a Target News without Modifying It, arXiv preprint arXiv:2312.15228, 2023.

[248] Zob. R. Kozik, M. Ficco, A. Pawlicka, M. Pawlicki, F. Palmieri, M. Choraś, When explainability turns into a threat: Using xAI to fool a fake news detection method, Computers & Security, 132, 103599, 2023.

zaawansowaną technologią może próbować wyłączyć „system alarmowy" poprzez atak na algorytm – należy więc przewidywać takie próby. Przykładowo, w przytoczonym wcześniej NATO-wskim scenariuszu utworzenia *hubu* analizującego media społecznościowe, zakłada się wykorzystanie XAI właśnie po to, by śledzić słabe punkty modeli i wskazywać potencjalne luki wykorzystywane przez przeciwnika[249]. Rozwiązaniem komplementarnym jest zachowanie czynników ludzkich w nadzorze nad systemem – człowiek może czasem dostrzec nienaturalny schemat, który zmylił maszynę. Ważne jest także wdrożenie niezależnego audytu i testów penetracyjnych modeli AI używanych w bezpieczeństwie – analogicznie jak testuje się zabezpieczenia systemów teleinformatycznych, testuje się odporność algorytmów na manipulacje. Takie działania zwiększają pewność, że integracja AI z instytucjami bezpieczeństwa faktycznie podnosi poziom ochrony, a nie stwarza nowych wektorów ataku.

Pełna integracja systemów wykrywania dezinformacji z aparatem bezpieczeństwa państwa wiąże się z koniecznością rozstrzygnięcia istotnych kwestii prawnych i etycznych. Z punktu widzenia prawa należy wypracować równowagę między skutecznym działaniem służb w infosferze a ochroną swobód obywatelskich, w tym wolności słowa i prywatności[250]. W wielu państwach demokratycznych uznaje się, że przeciwdziałanie dezinformacji powinno koncentrować się głównie na kampaniach wrogich podmiotów zagranicznych lub na treściach zagrażających życiu i bezpieczeństwu publicznemu między innymi w kontekście pandemii[251].

Prawo wciąż nie definiuje w sposób jednolity pojęcia „dezinformacja", co skutkuje różnicami w regulacjach[252]. Przykładowo, we Francji ustawa przeciw „manipulacjom informacyjnym" umożliwia szybkie blokowanie ewidentnych *fake news* w trakcie kampanii wyborczych. Z kolei w Niemczech nałożono na platformy obowiązek regularnego usuwania treści godzących w bezpieczeństwo publiczne[253]. W Polsce nie istnieje bezpośredni zapis

[249] Zob. B. Pinhasov, R. Lapid, R. Ohayon, M. Sipper, Y. Aperstein, Attacks on Deepfake Detectors, Afeka – The Academic College of Engineering in Tel Aviv & Ben-Gurion University of the Negev & DeepKeep, 2024.

[250] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107(6), 1753-1820.

[251] European Commission, Communication on Tackling COVID-19 Disinformation, 2020.

[252] European Commission, The Digital Services Act, 2022.

[253] European Commission, Communication on Tackling COVID-19 Disinformation, 2020.

kryminalizujący samo rozpowszechnianie fałszywych informacji, o ile nie wyczerpują one znamion innego przestępstwa, np. pomówienia[254]. Od strony ochrony danych osobowych (m.in. w UE pod reżimem RODO) wymagane są jasne procedury dotyczące monitorowania treści i gromadzenia informacji o użytkownikach[255]. Wobec podejrzeń nadużyć państwa demokratyczne stosują różne formy kontroli sądowej czy parlamentarnej nad działaniami służb[256]. Przykładowo w Wielkiej Brytanii ujawnienie istnienia *Counter Disinformation Unit* (CDU) wywołało publiczną debatę parlamentarną, zaś w USA trwa dyskusja nad zgodnością monitoringu w Internecie z pierwszą poprawką do Konstytucji *(First Amendment*)[257]. Aspekty etyczne koncentrują się wokół konieczności utrzymania zaufania społecznego. Działania przeciw dezinformacji mogą bowiem być postrzegane jako „cenzura" i mogą być nadużywane przez autorytarne reżimy[258]. Dlatego demokracje wprowadzają mechanizmy przejrzystości, np. publiczne listy fałszywych narracji (*EUvs-Disinfo*) i współpracę z niezależnymi podmiotami (*think-tanki*, uczelnie, *fact--checkerzy*)[259]. Istotne jest także zapobieganie uprzedzeniom w systemach AI – modele narażone są na przechylanie się w stronę określonych narracji, co wymaga weryfikacji treści zróżnicowanych ideologicznie[260]. W komunikacji publicznej zaleca się unikanie retoryki konfrontacyjnej („wojna z *fake news*") na rzecz edukacji i tzw. *pre-bunkingu*, polegającego na uprzedzaniu społeczeństwa o możliwych próbach manipulacji[261]. Współpraca międzynarodowa jest kluczowym elementem skutecznej integracji systemów detekcji dezinformacji. Dezinformacja nie respektuje granic państwowych, dlatego G7 uruchomiła *Rapid Response Mechanism* (RRM), zaś NATO i UE zacieśniły współdziałanie w ramach wymiany danych i koordynacji działań[262,263].

---

[254] KRRiT, Raport o dezinformacji i fake news, Warszawa, 2019.

[255] European Data Protection Board, Guidelines on the Processing of Personal Data under the GDPR, 2019.

[256] NATO Public Diplomacy Division – Monitoring Hybrid Threats, 2022.

[257] US Congress, Hearing on Social Media and Law Enforcement, Washington: US Government Printing Office, 2021.

[258] Council of Europe, Towards a Framework for Artificial Intelligence, 2022.

[259] RUSI, AI and Disinformation: Policy Briefing, 2023.

[260] Zob. Y. Mirsky, W. Lee, The Creation and Detection of Deepfakes: A Survey, ACM Computing Surveys, 54(1), 2021.

[261] Zob. R. Chesney, D. Citron, Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics, Foreign Affairs, 98(1), 147-155.

[262] NATO, NATO's Approach to Countering Disinformation, 2021.

[263] European Commission, The Digital Services Act, 2022.

Wspólny rozwój algorytmów, dzielenie się bazami zagrożeń czy raportowanie przez platformy (np. *Meta, Google*) o wykrytych sieciach dezinformacyjnych stanowi fundament takiego podejścia[264]. Organizacje międzynarodowe takie jak ONZ, Rada Europy, prowadzą prace nad standardami prawnymi zapobiegającymi nadużyciom AI w rozprzestrzenianiu fałszywych treści[265].

Podsumowując, skuteczna integracja systemów wykrywania dezinformacji z instytucjami bezpieczeństwa państwa wymaga wieloaspektowych rozwiązań: od klarownych regulacji prawnych i gwarancji etycznych po szeroko zakrojoną współpracę międzynarodową i edukację społeczeństwa. Taka strategia okazuje się niezbędna w dobie zagrożeń hybrydowych, w których polem konfrontacji jest coraz częściej infosfera.

6.4. Przykłady praktycznego wykorzystania narzędzi analizy audiowizualnej w bezpieczeństwie narodowym

Analiza audiowizualna – obejmująca automatyczną interpretację danych dźwiękowych i wizualnych – odgrywa coraz większą rolę w systemach bezpieczeństwa narodowego wielu państw. Zaawansowane algorytmy AI i techniki uczenia maszynowego potrafią w czasie rzeczywistym wyłuskiwać istotne informacje z nagrań audio/wideo, wspierając organy bezpieczeństwa w reagowaniu na zagrożenia. Poniżej przedstawiono konkretne scenariusze zastosowań takich narzędzi w różnych obszarach działań: wywiadowczych i operacyjnych, wojskowych, policyjnych, ratowniczych, antyterrorystycznych, a także w cyberbezpieczeństwie i analizie danych z *dark webu*. Omówiono również rzeczywiste przykłady wdrożeń tych technologii w różnych krajach (m.in. Chiny, USA, Izrael, państwa UE, Kanada), wskazując jednocześnie na ograniczenia, zagrożenia oraz proponowane rozwiązania, takie jak dostrajanie modeli (*fine-tuning*) czy współpraca człowieka z AI.

W środowisku wywiadowczym i operacyjnym narzędzia analizy audio-wideo służą do automatycznego pozyskiwania informacji z ogromnych zbiorów nagrań, które wcześniej musiały być żmudnie analizowane przez analityków. Przykładowo, agencje wywiadowcze wykorzystują

[264] Google Cloud, Locked Shields Case Study, 2023.
[265] Council of Europe, Towards a Framework for Artificial Intelligence, 2022.

rozpoznawanie mowy do transkrypcji i filtrowania przechwyconych rozmów telefonicznych czy nagrań audio, co umożliwia szybkie wychwycenie słów kluczowych związanych z działalnością przestępczą lub terrorystyczną. Zaawansowane systemy potrafią identyfikować głosy konkretnych osób (technika *voice fingerprinting*), co jest wykorzystywane do namierzania podejrzanych w tłumie sygnałów – np. rozpoznanie głosu lidera organizacji terrorystycznej podczas rozmowy może pozwolić na ustalenie jego lokalizacji[266]. W domenie wideo, wywiad często sięga po rozpoznawanie twarzy na zdjęciach i nagraniach z ukrytych kamer lub przechwyconych materiałów propagandowych. Pozwala to zidentyfikować uczestników tajnych spotkań czy obozów szkoleniowych na podstawie ujawnionych filmów. Przykładem może być analiza nagrań propagandowych tzw. Państwa Islamskiego – specjaliści, wspierani przez algorytmy rozpoznawania obrazu, potrafili wyodrębnić charakterystyczne elementy tła i na tej podstawie ustalić przybliżone miejsca nagrań, co pomogło w namierzeniu kryjówek terrorystów[267]. Innym scenariuszem wywiadowczym jest zastosowanie technik analizy audiowizualnej w OSINT (ang. *Open Source Intelligence*). Ogromna ilość treści w mediach społecznościowych i na jawnie dostępnych nagraniach wideo może zawierać cenne informacje. Automatyczna analiza umożliwia np. skanowanie publicznie dostępnych filmów z demonstracji czy zamieszek w celu identyfikacji prowodyrów i przebiegu incydentu. Wykorzystuje się do tego zarówno detekcję twarzy, sylwetki (identyfikacja po ubiorze, charakterystycznej posturze), jak i analizę dźwięku – np. rozpoznanie okrzyków tłumu wskazujących na zamiary eskalacji przemocy. Dzięki takim narzędziom służby mogą szybciej reagować na planowane akty agresji lub monitorować nastroje społeczne w sytuacjach kryzysowych. Na styku działań operacyjnych i wywiadowczych warto wspomnieć o zastosowaniu analizy audio-wideo w monitorowaniu *dark webu* pod kątem działalności przestępczej. Materiały multimedialne publikowane w ukrytych częściach Internetu – np. filmy z deklaracjami lub instruktażami terrorystycznymi – są automatycznie analizowane przez wyspecjalizowane algorytmy. Pozwala to

[266] Zob. L. Patino, J. Caicedo, R. Arguello, H. Gomez, Speaker recognition approach using mel-frequency cepstral coefficients and deep neural networks, IEEE Access, 8, 91877-91889.

[267] Zob. B. Pinhasov, R. Lapid, R. Ohayon, M. Sipper, Y. Aperstein, Attacks on Deepfake Detectors, Afeka – The Academic College of Engineering in Tel Aviv & Ben-Gurion University of the Negev & DeepKeep, 2024.

służbom wychwycić np. pojawienie się nowego nagrania lidera organizacji ekstremistycznej i natychmiast poddać je analizie (sprawdzenie autentyczności, wydobycie danych o lokalizacji z tła dźwiękowego lub obrazu). Takie rozwiązania są stosowane m.in. przez USA i ich sojuszników – *Five Eyes,* którzy dzielą się rezultatami automatycznej analizy przechwyconych materiałów audiowizualnych w ramach współpracy wywiadowczej[268].

Na współczesnym polu walki systemy analizy obrazu i dźwięku pełnią kluczową rolę w rozpoznaniu i świadomości sytuacyjnej. Wojskowe drony i samoloty rozpoznawcze rejestrują codziennie terabajty nagrań wideo – obrazy z kamer dziennych, noktowizyjnych czy termowizyjnych – których ręczne przejrzenie przekracza możliwości ludzkich analityków. Dlatego Departament Obrony USA uruchomił w 2017 r. inicjatywę *Project Maven* (oficjalnie *Algorithmic Warfare Cross-Functional Team*) w celu zastosowania uczenia głębokiego do automatycznej analizy materiałów wideo z dronów. Zadaniem *Maven* jest wykrywanie na nagraniach istotnych obiektów (np. pojazdów, osób, broni) i zdarzeń, tak aby analitycy mogli nawet trzykrotnie szybciej przetwarzać dane niż dotychczas. Podczas testów wykazano, że algorytmy mogą skutecznie identyfikować cele, jak konwoje wroga czy stanowiska artylerii, przy czym ostateczna decyzja należy do operatora. Piłkownik Drew Cukor, kierujący zespołem *Algorithmic Warfare Cross-Functional Team* w ramach programu *Project Maven*, podkreśla, że sztuczna inteligencja ma na celu wspieranie operatorów w analizie danych, a nie podejmowanie decyzji o wyborze celów bojowych: „*AI nie będzie samodzielnie wybierać celu… zadaniem AI jest uzupełnić pracę człowieka*"[269]. Projekt *Maven* był z powodzeniem używany w operacjach przeciwko ISIS i innych konfliktach po 2018 roku, stając się katalizatorem szerszego wdrażania AI w armii USA. Obecnie podobne rozwiązania rozwijają także Chiny i Rosja, obawiając się pozostania w tyle w tzw. wyścigu zbrojeń AI[270]. W zastosowaniach wojskowych analiza audiowizualna to nie tylko drony. Na poziomie taktycznym żołnierze korzystają z systemów wsparcia pola walki, które integrują analizę obrazu z wielu czujników. Przykładem są inteligentne gogle rzeczywistości

[268] Zob. Counter Terrorism Policing, Five Eyes Insights: Young People and Violent Extremism – A Call for Collective Action, 2024.

[269] Zob. C. Pellerin, Project Maven to Deploy Computer Algorithms to War Zone by Year's End, U.S. Department of Defense, 2017.

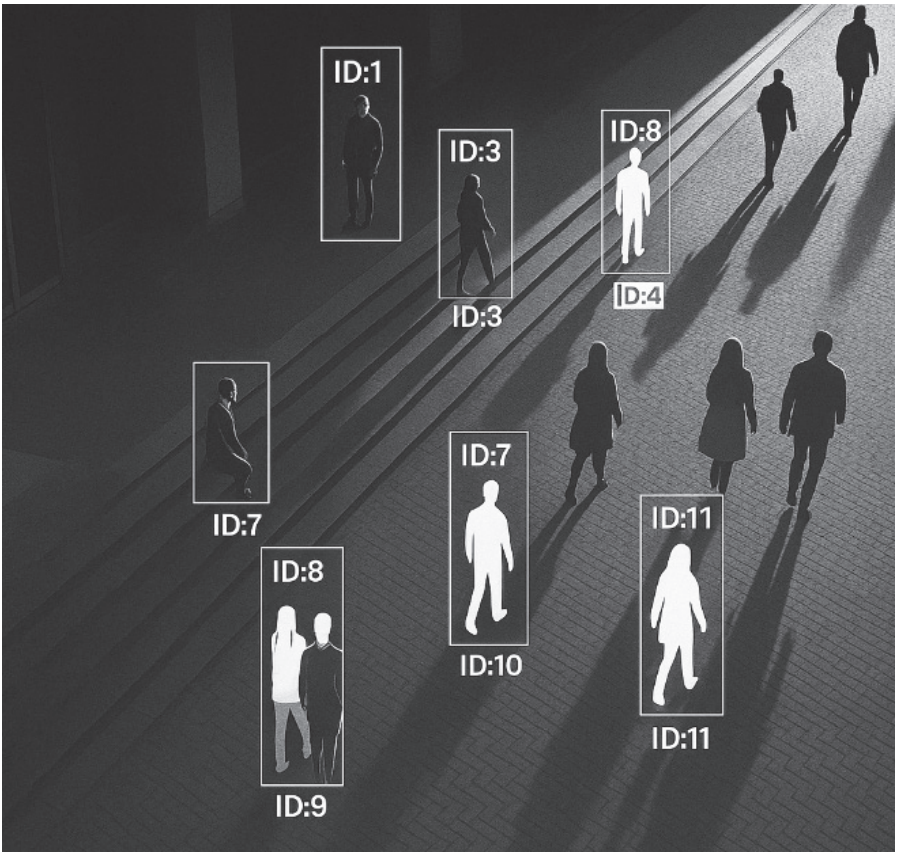
[270] Zob. RAND Corporation, Artificial Intelligence and National Security, RAND Corporation, 2019.

rozszerzonej IVAS (ang. *Integrated Visual Augmentation System)*, nad którymi pracuje armia USA. Gogle te, wyposażone w kamery i mikrofony, wykorzystują AI do oznaczania na bieżąco wykrytych przeciwników w polu widzenia żołnierza, rozpoznawania odgłosów (np. kierunku strzałów) i dostarczania powiadomień dźwiękowych. Pozwala to na szybszą reakcję – system podpowiada, skąd padł strzał albo gdzie przemieszcza się wróg, bazując na analizie audio (detekcja wystrzału) i wideo (śledzenie ruchu). Izrael, znany z zaawansowanych technologii wojskowych, również wdraża elementy analizy obrazu w swoich systemach bojowych. W pojazdach opancerzonych instalowane są kamery 360° sprzężone z algorytmami wykrywania zagrożeń, co ostrzega załogę przed zbliżającym się przeciwnikiem lub kierunkiem ataku. Z kolei systemy obronne na granicach np. wzdłuż Strefy Gazy, korzystają z automatycznej analizy obrazu do detekcji intruzów – wirtualne ogrodzenia rozpoznają sylwetkę człowieka przekraczającego linię graniczną i natychmiast alarmują posterunki. W ramach działań wojskowych istotne jest również rozpoznanie sygnałów dźwiękowych na polu walki. Akustyczne systemy wykrywania strzałów i eksplozji, wyposażone w AI, potrafią natychmiast zlokalizować źródło dźwięku wystrzału artyleryjskiego czy salwy karabinu. Przykładowo, amerykański system Boomerang wykorzystuje mikrofony rozstawione na pojazdach do identyfikacji kierunku, z którego padł strzał snajpera – analiza różnicy czasu dotarcia dźwięku do czujników jest wspomagana przez algorytm, który odfiltrowuje szum tła i klasyfikuje charakterystyczny dźwięk wystrzału. Podobnie systemy nasłuchu SONAR i RADAR z funkcjami AI na okrętach wykrywają odgłosy okrętów podwodnych czy dronów w powietrzu, wspierając bezpieczeństwo morskie i powietrzne. Warto zauważyć, że zaawansowane algorytmy analizy audiowizualnej stały się także elementem symbiozy człowieka i maszyny na polu walki. Operatorzy otrzymują wyniki analizy np. zaznaczone potencjalne cele na obrazie z drona, w formie sugestii, które mogą zweryfikować. Taki model *Human-AI teaming* sprawdza się – USA, Australia czy NATO podkreślają, że najlepsze efekty daje połączenie szybkości AI w wykrywaniu zagrożeń z intuicją i doświadczeniem człowieka[271]. Przykładem mogą być wspólne centra analiz koalicyjnych, gdzie oficerowie z różnych państw na bieżąco otrzymują alerty generowane automatycznie z strumieni danych

[271] National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. Washington, DC: The National Academies Press, 2021.

ISR (ang. *Intelligence, Surveillance, Reconnaissance*) i podejmują decyzje na podstawie zweryfikowanych już przez AI informacji.

Jednym z najbardziej rozpowszechnionych obszarów praktycznego wykorzystania analizy wideo i audio jest monitoring miejski i praca policji. W wielu metropoliach świata wdraża się inteligentne systemy analizy obrazu z kamer monitoringu miejskiego CCTV. Potrafią one automatycznie wykrywać podejrzane zdarzenia i osoby, odciążając operatorów centrów monitoringu. Przykładem funkcjonowania inteligentnego monitoringu miejskiego może być automatyczne rozpoznawanie nietypowych zachowań w przestrzeni publicznej dzięki zastosowaniu algorytmów sztucznej inteligencji przedstawione na rysunku 4.



Rysunek 4. Przykład działania inteligentnego systemu monitoringu miejskiego – automatyczne wykrywanie i śledzenie osób w przestrzeni publicznej z wykorzystaniem algorytmów sztucznej inteligencji

Źródło: opracowanie własne.

Zaznaczone na rysunku 4 ramki przedstawiają przykładowy sposób identyfikacji osób, które następnie mogą być automatycznie porównywane z istniejącymi bazami danych, np. osób poszukiwanych czy zaginionych. Algorytmy, wykorzystując zaawansowane techniki rozpoznawania twarzy oraz analizy sylwetek i zachowań, potrafią szybko wychwycić podejrzane osoby w tłumie i poinformować o tym odpowiednie służby. Takie rozwiązania pozwalają na błyskawiczną reakcję oraz skuteczne zwiększenie bezpieczeństwa w miejscach o dużym natężeniu ruchu, takich jak place, dworce kolejowe czy lotniska. Chiny są tu przykładem najbardziej rozbudowanego systemu nadzoru wizyjnego. Realizowany od 2015 r. program *Sharp Eyes* (chiń. *Xueliang*), będący elementem państwowej strategii „bezpieczne miasto", dążył do objęcia 100% przestrzeni publicznej monitoringiem do roku 2020. W ramach tego programu zainstalowano setki milionów kamer – szacunki mówiły o ponad 200 milionach kamer w całym kraju już w 2017 r. Kamery te są zintegrowane z systemami automatycznego rozpoznawania twarzy, co umożliwia natychmiastową identyfikację osób poszukiwanych w czasie rzeczywistym. Głośnym przykładem skuteczności, choć i kontrowersji, było ujęcie poszukiwanego przestępcy na koncercie w Chinach – kamery rozpoznały jego twarz w tłumie 60 tysięcy ludzi, prowadząc do aresztowania[272]. Chińskie systemy, jak *SkyNet* czy *Safe Cities*, integrują także analizę zachowań: algorytmy wychwytują bójki, gromadzenie się tłumu w nietypowym miejscu czy porzucone pakunki mogące wskazywać na zagrożenie bombowe. Co ciekawe, program *Sharp Eyes* angażował również obywateli – w niektórych regionach udostępniono mieszkańcom podgląd z wybranych kamer przez dedykowane aplikacje, zachęcając do zgłaszania incydentów. Taka masowa inwigilacja budzi jednak ogromne wątpliwości natury etyczno-prawnej. W krajach zachodnich podejście do monitoringu wspomaganego AI jest ostrożniejsze, ale również się rozwija. Stany Zjednoczone w wielu miastach wdrożyły systemy analizy nagrań z kamer ulicznych, choć nie na taką skalę jak Chiny. Przykładem jest *Project Green Light* w Detroit – sieć kamer HD na stacjach benzynowych i w sklepach, która w połączeniu z analizą wideo w czasie rzeczywistym umożliwia policji szybkie reagowanie na przestępstwa. Dodatkowo, wiele departamentów policji w USA korzysta z komercyjnych narzędzi do rozpoznawania twarzy,

[272] Zob. Z. Choińska, Wykorzystanie technologii rozpoznawania twarzy w Chińskiej Republice Ludowej, Folia Iuridica Universitatis Wratislaviensis, 11(1), 52-75.

takich jak *Clearview* AI. Aplikacja ta dysponuje bazą ponad 3 miliardów zdjęć zebranych z Internetu między innymi z mediów społecznościowych i pozwala na identyfikację osoby na zdjęciu poprzez dopasowanie do tej ogromnej bazy. Według dochodzenia *New York Times*, już w 2020 r. ponad 600 organów ścigania korzystało z *Clearview* AI w USA i innych krajach[273].

Choć narzędzie to znacząco przyspiesza identyfikację podejrzanych np. na podstawie zdjęć z miejsca przestępstwa, to jego użycie spotkało się z krytyką obrońców prywatności, a w Europie nawet z pozwami i karami finansowymi za naruszenie ochrony danych osobowych. W Europie rozwój inteligentnego monitoringu jest ograniczany przez surowe regulacje prywatności, ale mimo to prowadzone są pilotaże. Wielka Brytania testowała w latach 2016-2019 system rozpoznawania twarzy na żywo LFR (ang. *Live Facial Recognition*) w miejscach publicznych przez policję Metropolitalną w Londynie. Niezależny audyt wykazał jednak, że system ten działał zawodnie – poprawność identyfikacji wyniosła zaledwie ok. 19%, tylko co piąta wskazana osoba była faktycznie poprawnie zidentyfikowana, co oznaczało liczne fałszywe alarmy. W efekcie osoby niewinne bywały zatrzymywane omyłkowo, a eksperci uznali, że takie użycie może naruszać prawa człowieka[274]. Policja w Londynie wstrzymała więc szerokie wdrożenie i skupiła się na dopracowaniu algorytmów oraz procedur. Z kolei w Niemczech w 2018 r. przeprowadzono eksperyment na dworcu Berlin Südkreuz: testowano tam różne systemy rozpoznawania twarzy oraz algorytmy wykrywania pozostawionych bagaży i nietypowych zachowań. Ministerstwo spraw wewnętrznych podało, że przy łączeniu wyników trzech różnych systemów udało się osiągnąć 80% skuteczność identyfikacji osób . Jednak organizacja Chaos Computer Club wykazała, że pojedyncze systemy miały skuteczność znacznie niższą (ok. 68%), a wynik 80% uzyskano jedynie sumując detekcje wszystkich naraz[275]. Mimo to uznano, że technologia rokuje i planowano rozszerzenie monitoringu na 134 stacje kolejowe

[273] Zob. K. Hill, The Secretive Company That Might End Privacy as We Know It, in Ethics of Data and Analytics, 1st edn, Auerbach Publications, 2022, s. 1-8.

[274] Zob. P. Fussey, D. Murray, Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology, Human Rights, Big Data & Technology Project, University of Essex, 2023.

[275] Zob. A. V. Eireiner, Imminent dystopia? Media coverage of algorithmic surveillance at Berlin-Südkreuz, Internet Policy Review, 9(1), 2020.

i 14 lotnisk w Niemczech[276]. Ostatecznie jednak opór społeczny i obawy prawne zahamowały te plany, a nacisk położono na mniej inwazyjne rozwiązania, jak analiza wideo bez identyfikacji personalnej (np. detekcja zdarzeń). Wśród praktycznych policyjnych zastosowań analizy audiowizualnej warto wymienić systemy wykrywania strzałów w miastach. Technologia taka jak amerykański ShotSpotter wykorzystuje sieć mikrofonów rozmieszczonych na obszarach miejskich o wysokiej przestępczości. Gdy dojdzie do strzału z broni palnej, system automatycznie rozpoznaje charakterystyczny dźwięk wystrzału, lokalizuje go dzięki pomiarowi czasu dotarcia do kilku czujników i powiadamia najbliższy patrol policji w ciągu kilku sekund. Obecnie systemy *ShotSpotter* (lub ich odpowiedniki) działają w ponad 130 miastach USA, a także w niektórych miastach w Ameryce Południowej i RPA. To narzędzie znacząco skraca czas reakcji na strzelaniny i pozwala służbom dotrzeć na miejsce nawet wtedy, gdy nikt nie zdąży zadzwonić pod numer alarmowy. Wadą są jednak fałszywe alarmy – głośne fajerwerki lub trzask mogą czasem zostać błędnie zaklasyfikowane jako strzał, co skutkuje niepotrzebnym wysłaniem patroli. Innym przykładem są systemy do automatycznej analizy nagrań powypadkowych i powywiadowczych. Policje wielu krajów używają oprogramowania przyspieszającego przegląd tysięcy godzin nagrań z kamer przemysłowych podczas dochodzeń. Izraelska firma *BriefCam* opracowała technologię tzw. video synopsis, która kondensuje wiele godzin filmu w kilka minut, wyświetlając jednocześnie na ekranie wszystkie poruszające się obiekty z całego okresu nagrania. Pozwala to śledczemu szybko wychwycić interesujący moment m.in. pojawienie się podejrzanego samochodu, bez oglądania materiału w normalnym czasie. Takie rozwiązanie było wykorzystane m.in. przez policję w Bostonie po zamachu bombowym podczas maratonu w 2013 r., co ułatwiło zidentyfikowanie sprawców na podstawie zapisu z kamer. W kontekście bezpieczeństwa publicznego analizatory audio znajdują też zastosowanie w systemach alarmowych. Przykładowo, niektóre miasta w Holandii eksperymentowały z czujnikami akustycznymi na ulicach, które potrafią rozpoznać odgłosy tłuczonego szkła, krzyków czy kłótni – mogących sugerować bójkę lub akt wandalizmu. Po wykryciu takiego dźwięku, system automatycznie alertuje straż miejską lub policję, wskazując lokalizację zdarzenia. Choć technologia ta jest wciąż roz-

[276] Euractiv, Germany's plans for automatic facial recognition meet fierce criticism, Euractiv, 2020.

wijana, ma potencjał skrócić czas reakcji na incydenty losowe w przestrzeni publicznej.

Zapobieganie i zwalczanie terroryzmu stanowi kolejny obszar, w którym narzędzia analizy audiowizualnej są intensywnie wykorzystywane przez służby bezpieczeństwa państw. Kluczowe jest tutaj wczesne wykrywanie zagrożeń – identyfikacja osób planujących zamach lub rozpoznanie podejrzanego zachowania zanim dojdzie do ataku. W przestrzeniach publicznych narażonych na ataki jak lotniska, dworce, czy stadiony stosuje się kombinację analiz wideo i audio. Kamery na lotniskach coraz częściej wyposażane są w moduły AI, które skanują twarze podróżnych i porównują je z bazami danych osób poszukiwanych, np. międzynarodowymi listami terrorystów. W USA systemy takie działają pod egidą TSA na wybranych lotniskach, zaś w Europie trwają testy (np. lotnisko Schiphol w Amsterdamie eksperymentowało z rozpoznawaniem twarzy przy bramkach bezpieczeństwa). Ważnym aspektem jest tu precyzja – przy milionach pasażerów fałszywy alarm musi być zminimalizowany. W Izraelu, który od lat zmaga się z zagrożeniem terrorystycznym, stosuje się wielowarstwową kontrolę. Kamery z AI obserwują zachowanie osób jeszcze przed wejściem do terminala, wychwytując np. oznaki silnego zdenerwowania, nietypowe unikanie kamer czy podejrzane pakunki. System może generować ostrzeżenie typu „osoba podejrzanie się zachowująca w strefie X", co skutkuje dyskretnym sprawdzeniem danej osoby przez służby lotniskowe. Kluczowe znaczenie ma także analiza materiałów propagandowych i komunikacyjnych organizacji terrorystycznych. Służby wywiadowcze monitorują Internet (w tym *dark web*) w poszukiwaniu nagrań wideo publikowanych przez grupy ekstremistyczne. Każde nowe wystąpienie przywódcy i nagranie audio-wideo jest natychmiast poddawane analizie przez algorytmy. Rozpoznanie mowy pozwala szybko wydobyć treść przemowy i wychwycić ewentualne zaszyfrowane przekazy lub groźby, analiza obrazu może ujawnić metadane w tym datę, przybliżoną lokalizację z układu cieni czy roślinności, a porównanie twarzy potwierdza tożsamość mówcy. Szeroko opisywany był przypadek identyfikacji notorycznego terrorysty z ISIS o pseudonimie „*Jihadi John*" – analiza nagrań egzekucji, w tym porównanie głosu i budowy ciała ze znanymi danymi wywiadowczymi, przyczyniła się do ustalenia jego tożsamości. Podobnie, w 2020 r. irackie służby bezpieczeństwa przy wsparciu technologii rozpoznawania twarzy ujęły ukrywającego się lidera ISIS, gdy

ten pojawił się w krótkim filmie propagandowym. Algorytm automatycznie dopasował jego wizerunek do zdjęć z posiadanej przez służby bazy danych osób poszukiwanych[277]. Wspomniany przypadek dobrze ilustruje, jak narzędzia analizy audiowizualnej zwiększają możliwości operacyjne służb antyterrorystycznych, zwłaszcza w kontekście szybkiej identyfikacji i lokalizacji kluczowych członków organizacji terrorystycznych na podstawie minimalnej ilości dostępnych materiałów wizualnych. Izrael stanowi istotne studium przypadku intensywnego wykorzystania analizy audiowizualnej w działaniach antyterrorystycznych, zwłaszcza na terenach okupowanych. Około 2020 roku izraelska armia wdrożyła kontrowersyjny program o nazwie *Blue Wolf*, w ramach którego żołnierze wyposażeni w aplikacje mobilne rejestrowali zdjęcia napotkanych Palestyńczyków na Zachodnim Brzegu. Aplikacja natychmiast porównywała zeskanowaną twarz ze specjalną bazą danych, określaną przez izraelskich żołnierzy jako „tajny Facebook Palestyńczyków", po czym wyświetlała na ekranie smartfona kolorową sygnalizację (zielony, żółty lub czerwony), informując, czy daną osobę należy przepuścić, zatrzymać czy aresztować. Żołnierze rywalizowali nawet w zbieraniu zdjęć, co ilustruje skalę inwigilacji – w ciągu kilku miesięcy baza wzrosła o tysiące nowych rekordów. Równolegle, w miejscowościach takich jak Hebron, zainstalowano stacjonarne kamery wyposażone w funkcję automatycznego rozpoznawania twarzy, co pozwalało na identyfikację osób zbliżających się do punktów kontrolnych jeszcze zanim okazały dokumenty. System ten utworzył rozległą sieć nadzoru, umożliwiającą śledzenie ruchu osób uznawanych przez izraelskie służby za potencjalnych zamachowców lub osoby powiązane z organizacjami ekstremistycznymi. Izraelskie służby argumentują, że stosowanie tych technologii jest konieczne dla zapewnienia bezpieczeństwa obywateli przed atakami terrorystycznymi. Jednak organizacje broniące praw człowieka wskazują, że działania te mogą stanowić poważne naruszenie prywatności oraz być elementem systematycznego profilowania całej populacji palestyńskiej na terenach okupowanych. Analiza audiowizualna coraz częściej wspiera również kontrterrorystyczne jednostki specjalne podczas operacji bezpośrednich. Na przykład, przed szturmem na kryjówkę terrorystów, analitycy mogą wykorzystywać drony wyposażone

[277] Zob. S. Peterson, B.A. Jackson, D. Woods, E. Treyger, Finding a Broadly Practical Approach for Regulating the Use of Facial Recognition by Law Enforcement, RAND Corporation, 2023.

w kamery termowizyjne do obserwacji obiektu. Systemy sztucznej inteligencji automatycznie zliczają osoby obecne wewnątrz budynku na podstawie ich sygnatur cieplnych oraz sylwetek. Jednocześnie urządzenia nasłuchowe takie jak geofony są w stanie wychwycić rozmowy, ruch osób lub odgłosy manipulowania bronią, a specjalistyczne algorytmy klasyfikują te dźwięki, dostarczając dowódcom precyzyjnych informacji dotyczących sytuacji operacyjnej. W trakcie samego szturmu kamery zamontowane na hełmach operatorów mogą przesyłać strumień obrazu do centrum dowodzenia, gdzie systemy analityczne w czasie rzeczywistym wspierają dowodzących poprzez automatyczne oznaczanie i śledzenie pozycji zakładników oraz napastników. Ponadto, warto zwrócić uwagę na wyspecjalizowane systemy akustyczne stosowane do ochrony infrastruktury krytycznej przed potencjalnymi atakami terrorystycznymi. Przykładem są wybrane budynki rządowe w USA czy Izraelu są wyposażone w czujniki ultradźwiękowe i sejsmiczne, które umożliwiają wykrywanie prób podkopów lub ingerencji w konstrukcję budynku między innymi przy próbach umieszczenia ładunków wybuchowych w fundamentach. Inteligentna analiza sygnałów akustycznych pozwala skutecznie odróżnić normalne drgania miejskie czy odgłosy infrastruktury od celowego działania ludzkiego, automatycznie alarmując ochronę obiektu w przypadku wykrycia podejrzanej aktywności.

Narzędzia analizy audio-wideo okazują się niezwykle pomocne nie tylko w działaniach przeciwdziałających zagrożeniom stworzonym przez człowieka, ale także w sytuacjach kryzysowych spowodowanych klęskami żywiołowymi czy wypadkami. Służby ratownicze coraz częściej korzystają z dronów, kamer oraz mikrofonów nasobnych wspomaganych AI, aby szybciej lokalizować ofiary i podejmować właściwe działania. W przypadku klęsk żywiołowych, takich jak trzęsienia ziemi czy huragany, kluczowe jest szybkie zidentyfikowanie osób uwięzionych pod gruzami lub odciętych od pomocy. Tradycyjnie ekipy ratunkowe używają psów tropiących i nasłuchują wołań o pomoc – obecnie wspiera ich technologia. Drony z kamerami wysokiej rozdzielczości oraz termowizyjnymi mogą przeszukiwać rozległy teren znacznie szybciej niż ludzie. Algorytmy analizy obrazu na żywo rozpoznają kształt ludzkiego ciała lub jego fragmentów wśród rumowisk, a także ślady ciepła ludzkiego organizmu na obrazie termicznym. Pozwala to skierować ratowników dokładnie tam, gdzie znajdują się poszkodowani. W 2017 r. po trzęsieniu ziemi w Meksyku z powodzeniem wykorzystano drony AI

– przelatując nad zawalonymi budynkami, ich systemy wykryły sylwetki dwóch ocalałych osób machających z okna na wyższej kondygnacji, które były niewidoczne z poziomu ulicy[278]. Wcześniej mogłyby one zostać odnalezione dopiero po wielu godzinach. Analiza dźwięku jest równie istotna w akcjach ratunkowych. W sytuacjach, gdy ludzie uwięzieni pod gruzami nie są widoczni, często wołają o pomoc lub wydają odgłosy poprzez stukanie w rury, czy ściany. Specjalistyczne mikrofony kierunkowe i geofony sprzężone z AI potrafią wychwycić nawet słaby dźwięk wołania i odróżnić go od odgłosów maszyn ratowniczych czy wiatru. Algorytm wzmacnia istotne sygnały i filtruje szum, po czym klasyfikuje dźwięk oceniając, czy jest to głos ludzki. Co więcej, poprzez analizę czasu dotarcia dźwięku do kilku rozmieszczonych czujników, system trianguluje położenie poszkodowanego. Takie systemy testowano z sukcesem we Włoszech po zawaleniu mostu Morandi w Genui w roku 2018, co pomogło zlokalizować ostatniego ocalałego dzięki ledwo słyszalnym stukom wydobywającym się spod ciężkich płyt. Służby ratownicze korzystają również z analizy audiowizualnej przy koordynacji akcji poszukiwawczych z udziałem wielu zespołów. W centrach zarządzania kryzysowego spływa ogromna liczba informacji: zgłoszenia telefoniczne, nagrania z kamer miejskich, zdjęcia nadsyłane przez obywateli. Inteligentne systemy mogą automatycznie analizować treść tych zgłoszeń – na przykład infolinię 112 w niektórych krajach UE wspiera prototyp AI, który analizuje głos zgłaszającego pod kątem stresu i emocji, aby priorytetyzować najpoważniejsze przypadki. W straży pożarnej opracowano z kolei system, który nasłuchuje detektorów czadu i dymu rejestrowanych przez mikrofony miejskie – wykrycie charakterystycznego pisku alarmu przeciwpożarowego skutkuje automatycznym zgłoszeniem potencjalnego pożaru zanim jeszcze ktoś zadzwoni po pomoc. W terenowych akcjach poszukiwawczych z użyciem kamer AI potrafi analizować obraz z kamer na bieżąco i wypatrywać kolorów lub kształtów pasujących do poszukiwanego. Jeśli wiadomo, że zaginiony ma określony kolor kurtki, dron z kamerą przeszukujący las może automatycznie oznaczać na podglądzie wszelkie obszary z tym kolorem i alarmować operatora, gdy wykryje sylwetkę człowieka. Takie systemy testowano w Kanadzie podczas poszukiwania zaginionych

[278] Zob. S.M.S.M. Daud, M.Y.P.M. Yusof, C.C. Heo, L.S. Khoo, M.K.C. Singh, M.S. Mahmood, H. Nawawi, Applications of drone in disaster management: A scoping review, Science & Justice, 61(6), 674-685.

turystów w parkach – okazało się, że AI zauważyła człowieka w gęstym lesie na nagraniu drona o znalezienie kilkanaście minut wcześniej niż dostrzegłby go ludzki operator W efekcie osoba została szybciej odnaleziona i uratowana przed nocą w niskiej temperaturze. Również ratownictwo morskie korzysta z analizatorów wideo. Kamery monitorujące plaże i wybrzeże, wyposażone w AI, potrafią wykryć tonącą osobę po nieregularnych ruchach i zanikaniu sylwetki w wodzie, co wyzwala alarm i uruchomienie procedury ratunkowej zanim ktoś zareaguje. W Hiszpanii prowadzone są pilotaże takiego systemu na kąpieliskach morskich – wstępne wyniki sugerują, że może on skrócić czas zauważenia topiącego się pływaka o cenne sekundy.

Mimo imponujących możliwości, narzędzia analizy audio-wideo napotykają szereg ograniczeń i niosą ze sobą istotne zagrożenia. Fałszywe alarmy (ang. *false positives*) oraz przeoczenia (ang. *false negatives*) stanowią jedno z głównych wyzwań. Jak wskazują doświadczenia z systemami rozpoznawania twarzy, ich skuteczność bywa zawodna. Fałszywe alarmy obciążają służby (reakcje na zdarzenia, które nie miały miejsca lub dotyczą niewinnych osób), a przeoczenia mogą oznaczać niezauważenie realnego zagrożenia. Przyczyny bywają techniczne takie jak niewyraźny obraz, czy szum dźwięku lub mogą być związane z algorytmem – stronniczość danych treningowych. Z tym wiąże się problem *biasu* algorytmicznego – uprzedzeń zakodowanych w systemach AI. Badanie[279] wykazało, że komercyjne systemy rozpoznawania twarzy osiągały znacznie wyższą skuteczność dla twarzy osób rasy kaukaskiej niż dla osób o ciemniejszym odcieniu skóry. Błędy w rozpoznaniu dotyczyły zwłaszcza kobiet o ciemnej karnacji, gdzie wskaźnik błędnej klasyfikacji płci sięgał kilkudziesięciu procent, podczas gdy dla mężczyzn rasy białej wynosił poniżej 1%. Tego rodzaju nierówności przekładają się na ryzyko dyskryminacji – np. niewinnie idący ulicą Afroamerykanin może zostać błędnie zidentyfikowany jako poszukiwany przestępca, co prowadzi do nieuzasadnionego zatrzymania. Takie przypadki już miały miejsce w USA, gdzie co najmniej kilku obywateli zostało niesłusznie aresztowanych na skutek pomyłki systemu rozpoznawania twarzy. Rodzi to poważne konsekwencje prawne i społeczne, podważając zaufanie do technologii. Kolejnym zagrożeniem są ataki *adversarialne*, czyli celowe działania zmierzające do zmylenia systemów AI. Okazuje się, że algorytmy widzenia

[279] Zob. J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of Machine Learning Research, 81, 1-15.

komputerowego można oszukać poprzez wprowadzenie drobnych, pozornie nieistotnych modyfikacji obrazu lub otoczenia. Przykładowo, badacze z *Carnegie Mellon University* opracowali w 2016 r. metodę wydrukowania wzoru na oprawkach okularów, który zmyla system rozpoznawania twarzy – osoba mająca takie „inteligentne" okulary może zostać niewykryta albo nawet błędnie rozpoznana jako inna. Innymi słowy, prosta fizyczna manipulacja taka jak kolorowy wzór na okularach, zmieniła odbiór obrazu przez maszynę, podczas gdy człowiek nadal bez trudu rozpozna twarz. Podobne sztuczki działają w przypadku analizy dźwięku – tzw. *adversarial examples* w audio to specjalnie spreparowane szumy lub frazy, które dla człowieka brzmią jak zwykły szum tła, ale system rozpoznawania mowy interpretuje je jako konkretną komendę. Tego typu ataki mogą posłużyć przestępcom do obejścia zabezpieczeń. Prywatność i prawa obywatelskie stanowią kolejne pole do dyskusji. Wszechobecny monitoring wizyjny budzi zrozumiałe obawy o inwigilację społeczeństwa. W demokratycznych państwach pojawiają się pytania: gdzie jest granica między zapewnieniem bezpieczeństwa a naruszeniem prywatności obywateli? Gdy technologia umożliwia ciągłe śledzenie i rozpoznawanie każdego przechodnia na ulicy, istnieje ryzyko nadużyć i powstania społeczeństwa nadzorowanego. Przykłady z Chin czy wspomnianego izraelskiego programu *Blue Wolf* pokazują, że technologia może zostać użyta do systemowej dyskryminacji lub represji określonych grup. W krajach UE wykorzystanie analizy biometrycznej jak rozpoznawanie twarzy jest ograniczane przepisami – RODO traktuje dane biometryczne jako szczególnie wrażliwe. Błędy i odpowiedzialność prawna to kolejny aspekt. Jeżeli algorytm wskaże niewłaściwą osobę jako terrorystę, kto ponosi odpowiedzialność za ewentualne szkody np. niesłuszne zatrzymanie, czy naruszenie dóbr osobistych? Obecnie systemy AI są traktowane jako narzędzia wspomagające, a decyzje formalnie podejmują ludzie. Literatura przedmiotu wskazuje na efekt *automation bias* – skłonność operatorów do ufania sugestiom systemu nawet wbrew własnej intuicji[280]. Może to prowadzić do sytuacji, w której policjant uzna, że system „wie lepiej" i dokona zatrzymania, choć sam nie dostrzegł powodu – a potem okaże się, że był to błąd systemu. Nie sposób pominąć też kwestii odporności systemów analizy audiowizualnej na zakłócenia i sabotaż. Przeciwnik świadomy

[280] R. Parasuraman, D. H. Manzey, Complacency and Bias in Human Use of Automation, Human Factors, 2010.

wykorzystywania AI przez służby może próbować oślepić te systemy – np. stosując lasery do uszkadzania sensorów kamer monitoringu lub emitując silny hałas akustyczny, by zagłuszyć mikrofony (tzw. atak *denial-of-service* na system audio). Dla zaawansowanych grup przestępczych celem może być również zhakowanie takich systemów. Jeśli cyberprzestępcy uzyskają dostęp do miejskiego systemu CCTV z AI, mogą np. manipulować obrazem, wstawiając fałszywe obiekty, usuwając prawdziwe w nagraniach i zmylić służby.

W obliczu powyższych wyzwań trwają prace nad udoskonaleniem technologii oraz ustanowieniem ram regulacyjnych dla bezpiecznego i etycznego wykorzystania analizy audiowizualnej. Można wskazać kilka głównych kierunków rozwoju i strategii mitigacji zagrożeń:

- Dostosowywanie i doskonalenie modeli (ang. *fine-tuning*): Aby zmniejszyć liczbę błędów i uprzedzeń, modele AI powinny być regularnie trenowane na zróżnicowanych, aktualnych danych. *Fine-tuning* modeli do warunków lokalnych (np. uwzględnienie w zbiorze treningowym wizerunków mniejszości etnicznych żyjących w danym kraju) pomaga poprawić ich skuteczność i sprawiedliwość. Firmy technologiczne wdrażają procedury audytu algorytmów pod kątem *biasu* – np. sprawdzając, czy system rozpoznawania twarzy ma podobny poziom skuteczności dla różnych grup demograficznych i korygując go, jeśli tak nie jest. W przypadku systemów wykrywania zdarzeń, trening na danych z danego miasta, a nie tylko symulacyjnych, pozwala lepiej odróżnić zwyczajne sytuacje od nietypowych. Przykładowo, algorytm analizujący ruch uliczny może być dostrojony tak, by nie zgłaszać alarmu na widok tłumu przed stadionem w dniu meczu – rozpoznając, że jest to normalne w danej lokalizacji, w odróżnieniu od zgromadzenia tłumu w miejscu, gdzie zwykle jest pusto.
- Łączenie wielu źródeł (ang. *sensor fusion*): Połączenie analizy audio i wideo oraz innych czujników może zwiększyć wiarygodność wykrywania i zmniejszyć fałszywe alarmy. Jeśli system CCTV wykryje pozostawioną torbę na dworcu, może równocześnie nasłuchiwać odgłosów paniki lub krzyków – brak takich odgłosów może obniżyć priorytet alarmu, gdyż być może to tylko zagubiony bagaż, podczas gdy ich obecność go podwyższy. Wykorzystanie wielu algorytmów jednocześnie – np. trzech niezależnych systemów rozpoznawania twarzy zamiast jednego – i uzgadnianie ich wyników (ang. *ensemble*

learning) też podnosi dokładność. Niemcy testując kilka systemów jednocześnie uzyskali sumarycznie wyższą skuteczność, choć kosztem większej złożoności.

- Udział człowieka w pętli decyzyjnej (ang. *Human-AI teaming*): Konsensus wśród ekspertów bezpieczeństwa jest taki, że AI powinna wspierać, a nie zastępować funkcjonariuszy. Oznacza to utrzymanie człowieka „w pętli" – każda automatyczna detekcja powinna zostać zweryfikowana przez operatora zanim wywoła poważne skutki (np. zatrzymanie osoby). Praktyka ta jest wdrażana miedzy innymi w Londynie, gdzie po porażkach testów LFR obecnie planuje się używać rozpoznawania twarzy jedynie jako wskazówki: system wskaże potencjalne trafienie, ale decyzję o podjęciu interwencji podejmie funkcjonariusz porównując zdjęcie z bazy i obraz z kamery. Podobnie w projekcie *Maven* w wojsku USA założono, że AI podaje propozycję oznaczenia celu, ale człowiek ostatecznie autoryzuje każdy atak. Taki podział ról pozwala wykorzystać szybkość i spostrzegawczość AI, jednocześnie korzystając z rozwagi i kontekstu rozumianego przez człowieka. Ważne jest też szkolenie operatorów, by rozumieli możliwości i ograniczenia systemów – wtedy łatwiej wyłapią ewentualne błędy (np. wiedząc, że algorytm ma skłonność do fałszywych trafień w określonych warunkach oświetleniowych).
- Uodparnianie systemów na ataki (ang. *robust* AI): Trwają intensywne prace badawcze nad metodami zabezpieczania modeli przed atakami *adversarialnymi*. Jedną z technik jest *adversarial training*, czyli włączenie do treningu sieci neuronowej także przykładów zakłóconych *adversarialnie* – tak aby „nauczyła się" je ignorować. Inne podejście to wykrywanie, gdy dane wejściowe noszą cechy ataku – np. filtrowanie obrazu by usunąć podejrzane zakłócenia wzorca. W przypadku systemów fizycznych, inżynieryjnie zabezpiecza się sprzęt: kamery mogą mieć filtry i osłony utrudniające oślepienie laserem, mikrofony mogą być kalibrowane tak, by nie reagować na częstotliwości poza ludzkim głosem (co utrudnia ukrycie komendy w ultradźwiękach). Choć wyścig między atakującymi a obrońcami trwa, zwiększenie odporności systemów AI jest priorytetem przy ich wdrażaniu w krytycznych zastosowaniach.

- Regulacje prawne i standardy etyczne: W wielu krajach trwają prace nad uregulowaniem stosowania AI w bezpieczeństwie. Unia Europejska proceduje obecnie tzw. *AI Act*, który przewiduje zaliczenie systemów zdalnej identyfikacji biometrycznej do kategorii „wysokiego ryzyka" i nakłada na nie rygorystyczne wymogi (transparentność, audyt, nadzór człowieka). Już teraz w Europie obowiązują przepisy ograniczające przechowywanie danych z monitoringu i wymagające ich szyfrowania[281]. Ponadto, organizacje międzynarodowe jak INTERPOL i Europol publikują wytyczne etycznego użycia AI – np. raport Europolu z 2024[282] r. zaleca, by wdrażać analizę video stopniowo, z udziałem społeczeństwa, tylko do jasno zdefiniowanych celów i za zgodą odpowiednich szczebli dowodzenia. Podkreśla się, że policja powinna informować obywateli o używaniu takich technologii, aby budować zaufanie i akceptację społeczną (Borak, 2025). Ważne jest też prowadzenie ocen skutków dla prywatności (ang. *Privacy Impact Assessment*)[283] przed wdrożeniem nowych systemów – tak, by zawczasu zidentyfikować i zminimalizować ryzyka.
- Transparentność i rozliczalność: Proponuje się wprowadzenie mechanizmów logowania decyzji AI – system powinien zapisywać, jakie przesłanki (np. cechy obrazu) spowodowały daną alarmową klasyfikację. To ułatwi późniejszą kontrolę, dlaczego np. doszło do pomyłki. Rozwijana jest dziedzina XAI (ang. *Explainable AI)*, aby nawet złożone modele głębokie dały się częściowo objaśnić. W kontekście sądowym dyskutuje się, na ile dowody wygenerowane przez AI (np. „system wskazał tę osobę jako sprawcę na nagraniu") mogą być wykorzystane – konieczne mogą być niezależne ekspertyzy biegłych potwierdzające wiarygodność takiego wyniku.
- Współpraca międzynarodowa i wymiana danych: Zagrożenia bezpieczeństwa mają charakter transnarodowy, dlatego państwa współpracują w rozwijaniu i udoskonalaniu narzędzi analitycznych. Inicjatywy takie jak *Interpol Innovation Centre* czy wspólne ćwiczenia NATO

[281] Europejska Rada Ochrony Danych (EROD), Wytyczne 3/2019 w sprawie przetwarzania danych osobowych przez urządzenia wideo, EROD, 2019.

[282] Europol, AI and Policing, Europol, 2024

[283] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689, w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji, Dziennik Urzędowy UE, 2024.

w zakresie AI pozwalają na dzielenie się wiedzą o najlepszych praktykach[284]. Jednocześnie pojawia się postulat, by demokracje wypracowały wspólne standardy użycia tej technologii zgodne z wartościami praw człowieka, co utrudni eksport i nadużywanie inwigilacyjnych systemów przez reżimy autorytarne.

Podsumowując, narzędzia analizy audiowizualnej już teraz dostarczają służbom bezpieczeństwa narodowego nieocenionego wsparcia, przyspieszając wykrywanie zagrożeń i dostarczając informacji do działania. Ich praktyczne wykorzystanie obejmuje szerokie spektrum – od patrolowania ulic i granic, poprzez pole walki, po sferę cyberprzestrzeni. Rzeczywiste wdrożenia w różnych krajach pokazują zarówno potencjał, jak i pułapki tych rozwiązań. W nadchodzących latach zapewne będziemy świadkami dalszej ekspansji tych technologii, ale również zaostrzenia debaty nad tym, jak korzystać z nich odpowiedzialnie. Kluczem będzie znalezienie równowagi między bezpieczeństwem a prawami obywateli – tak, aby nowoczesne systemy analizy audio-wideo służyły ochronie społeczeństwa, a nie jego nadmiernej kontroli.

[284] Europol, AI and Policing, Europol, 2024.

VII

Badania empiryczne – skuteczność metod analizy treści audiowizualnych

7.1. Opis zastosowanej metodologii badawczej

Badania empiryczne, które przeprowadzono w niniejszej monografii, oparto na kompleksowej i interdyscyplinarnej metodologii badawczej, która obejmowała zarówno jakościowe, jak i ilościowe techniki analizy. Taki dobór metodologii wynikał ze specyfiki problemu badawczego, który dotyczy zjawiska manipulacji audiowizualnej oraz oceny skuteczności narzędzi umożliwiających ich identyfikację w kontekście bezpieczeństwa państwa. Zastosowanie mieszanych metod badawczych umożliwiło uzyskanie wiarygodnych, naukowo uzasadnionych oraz praktycznie użytecznych wyników.

Pierwszym krokiem realizacji procesu badawczego było opracowanie szczegółowego projektu metodologicznego, który zakładał etapową realizację badań oraz precyzyjne określenie wykorzystanych technik analitycznych. Na tym etapie przeprowadzono przegląd literatury naukowej oraz raportów eksperckich dotyczących zjawiska manipulacji audiowizualnej, ze szczególnym uwzględnieniem współczesnych technik deepfake oraz klasycznych metod cyfrowej dezinformacji audiowizualnej. To podejście pozwoliło na pogłębione zrozumienie charakteru zagrożeń oraz wyznaczenie głównych kierunków i kryteriów analizy empirycznej.

Badania empiryczne zrealizowano w dwóch powiązanych ze sobą, lecz wyraźnie odrębnych fazach. Pierwsza faza obejmowała jakościowe studium przypadków, skoncentrowane na pogłębionej analizie rzeczywistych przykładów manipulacji audiowizualnych, których wybór został dokonany celowo w oparciu o wcześniej ustalone kryteria istotności dla bezpieczeństwa

informacyjnego państwa. Analiza jakościowa polegała na dokładnej, ręcznej ocenie materiałów audiowizualnych przez eksperta badawczego, co pozwoliło na identyfikację charakterystycznych cech i mechanizmów manipulacji. W tej części badań istotna była analiza treści narracyjnych, wizualnych oraz dźwiękowych, umożliwiająca precyzyjne zdefiniowanie specyficznych technik stosowanych przez autorów manipulacji. Jakościowa ocena treści miała na celu nie tylko identyfikację metod dezinformacyjnych, ale także określenie ich możliwych skutków społecznych i politycznych w kontekście bezpieczeństwa państwa. Drugą, główną fazą badań były testy eksperymentalne, prowadzone w formacie ilościowym, których celem było zweryfikowanie skuteczności zaawansowanych technologii i algorytmów służących do wykrywania manipulacji audiowizualnych. W ramach tej części badań przygotowano odpowiednio dobraną próbę materiałów testowych, składającą się zarówno z materiałów autentycznych, które pełniły funkcję grupy kontrolnej, jak i ze zmanipulowanych materiałów, reprezentujących różne typy fałszerstw medialnych. Dobór tych materiałów został przeprowadzony zgodnie ze standardami naukowymi, w oparciu o szeroko stosowane w środowisku akademickim zbiory danych referencyjnych, takie jak FaceForensics++ oraz DeepFake Detection Challenge. Taki dobór gwarantował reprezentatywność próby oraz umożliwiał późniejszą porównywalność wyników z rezultatami uzyskanymi przez inne zespoły badawcze na świecie.

W toku eksperymentów zastosowano narzędzia analityczne oparte przede wszystkim na uczeniu maszynowym oraz głębokich sieciach neuronowych. Wykorzystano specjalistyczne algorytmy analizy obrazu bazujące na konwolucyjnych sieciach neuronowych (CNN), szczególnie skuteczne w identyfikacji subtelnych, trudno dostrzegalnych przez człowieka artefaktów cyfrowych. Ponadto, do analizy treści audio wykorzystano rekurencyjne sieci neuronowe (RNN), a w szczególności modele oparte na komórkach pamięci długoterminowej (LSTM), umożliwiające wychwycenie niezgodności w strukturze sygnałów dźwiękowych, charakterystycznych dla cyfrowo generowanych wypowiedzi. Uzupełnieniem tych zaawansowanych metod była analiza forensyczna, obejmująca sprawdzanie spójności metadanych, jakości cyfrowej kompresji oraz synchronizacji warstwy wizualnej i dźwiękowej, co stanowi klasyczną technikę weryfikacji integralności materiałów medialnych. Przed przystąpieniem do właściwych testów eksperymentalnych wszystkie narzędzia analityczne zostały skalibrowane oraz poddane

walidacji wstępnej, przeprowadzonej na referencyjnych zestawach danych testowych. Ten etap był niezbędny, aby upewnić się, że algorytmy działają poprawnie i osiągają wyniki porównywalne z deklarowanymi w literaturze przedmiotu. W samych eksperymentach zastosowano metodę ślepej próby, gdzie klasyfikacja materiałów została przeprowadzona bez uprzedniej wiedzy o ich statusie (autentyczny/zmanipulowany), co zapewniło obiektywność i wiarygodność wyników. Każdy materiał audiowizualny był analizowany przez kilka niezależnych algorytmów, a wyniki były następnie agregowane w celu uzyskania finalnej klasyfikacji wraz z określeniem poziomu ufności. Uzyskane w eksperymentach ilościowych wyniki zostały poddane szczegółowej analizie statystycznej, w której uwzględniono kluczowe metryki jakości klasyfikacji – dokładność, czułość, specyficzność oraz poziomy błędów klasyfikacyjnych (fałszywe alarmy i przeoczenia). Zastosowanie analizy statystycznej umożliwiło obiektywne porównanie skuteczności poszczególnych metod oraz identyfikację ich słabych i mocnych stron.

Przedstawiona metodologia została tak skonstruowana, aby umożliwić uzyskanie wyczerpujących, obiektywnych oraz wiarygodnych wyników, które mogłyby stanowić solidną podstawę do dalszych analiz oraz do sformułowania rekomendacji praktycznych dotyczących zastosowania metod wykrywania manipulacji audiowizualnych w strategiach bezpieczeństwa państwa. W szczególności celem badań było precyzyjne zdefiniowanie skuteczności istniejących algorytmów analitycznych oraz określenie ich praktycznej użyteczności w realnych warunkach informacyjnych, co stanowi kluczowy element wzmacniania odporności państwa na współczesne zagrożenia dezinformacyjne.

7.2. Eksperymenty na zbiorach danych audiowizualnych – testowanie skuteczności algorytmów

W ramach badań empirycznych przeprowadzonych na potrzeby niniejszej monografii dokonano szeroko zakrojonych eksperymentów mających na celu ocenę skuteczności wybranych algorytmów do wykrywania manipulacji w rzeczywistych materiałach audiowizualnych. Badania zostały zrealizowane zgodnie z przyjętą metodologią i obejmowały trzy zasadnicze etapy: dobór i charakterystykę materiałów testowych, szczegółową analizę użytych

algorytmów oraz przeprowadzenie eksperymentów, które pozwoliły zweryfikować skuteczność tych metod w warunkach zbliżonych do operacyjnych. Na początkowym etapie badań dokonano precyzyjnego wyboru i identyfikacji materiałów audiowizualnych podlegających analizie. Kryteria selekcji opierały się na reprezentatywności, autentyczności oraz potencjalnym oddziaływaniu wybranych przypadków na percepcję społeczną oraz bezpieczeństwo informacyjne państwa. W ramach eksperymentów wykorzystano między innymi materiał audiowizualny z 2019 roku, który przedstawiał zmanipulowane nagranie amerykańskiej polityk Nancy Pelosi. Nagranie to, będące przykładem stosunkowo prostej manipulacji cyfrowej (tzw. *cheapfake*), opierało się na spowolnieniu wypowiedzi w taki sposób, aby wywołać wrażenie nietrzeźwości, co miało bezpośredni wpływ na wizerunek i postrzeganie polityczne przedstawionej osoby. Kolejnym przykładem wykorzystanym w eksperymentach było zmanipulowane nagranie audio z 2020 roku przedstawiające fałszywe oświadczenie Elona Muska na temat rzekomych planów inwestycyjnych. Pomimo szybkiego zdemaskowania tego fałszerstwa, nagranie przez krótki okres wywołało reakcje społeczne i biznesowe, uwypuklając tym samym potencjalne zagrożenia wynikające ze stosowania cyfrowych manipulacji audiowizualnych w przestrzeni publicznej. Kolejną częścią eksperymentów była szczegółowa charakterystyka narzędzi i algorytmów wykorzystywanych do analizy zebranych materiałów audiowizualnych. W ramach prowadzonych badań użyto zaawansowanych rozwiązań technologicznych opartych na sztucznej inteligencji, przede wszystkim konwolucyjnych sieci neuronowych (CNN), które są szczególnie skuteczne w wykrywaniu manipulacji wizualnych. Algorytmy te, bazujące na architekturze typu *Xception*, umożliwiły szczegółową identyfikację subtelnych artefaktów cyfrowych fałszerstw, takich jak nienaturalne ruchy twarzy, błędy w synchronizacji ruchu ust z warstwą audio, czy niezgodności związane z oświetleniem i teksturą obrazu. Równolegle zastosowano rekurencyjne sieci neuronowe (RNN), a dokładniej ich wariant LSTM, które umożliwiły analizę warstwy dźwiękowej materiałów audiowizualnych, identyfikując specyficzne anomalie sygnałów głosowych wynikające z cyfrowej syntezy mowy. Oprócz narzędzi opartych na technologiach sztucznej inteligencji, w badaniach uwzględniono również klasyczne metody cyfrowej analizy śledczej (*digital forensics*), obejmujące m.in. ocenę integralności metadanych, detekcję artefaktów powstałych podczas cyfrowej obróbki obrazu czy analizę niespójności warstw dźwiękowej i wizualnej materiałów audiowizualnych. Trzecim

etapem eksperymentalnym było przeprowadzenie szczegółowych testów skuteczności wybranych algorytmów. Każdy z przygotowanych wcześniej materiałów audiowizualnych został przetworzony przez zestaw wybranych narzędzi detekcyjnych, które następnie klasyfikowały nagrania jako autentyczne bądź zmanipulowane. Wyniki analiz uwzględniały także poziom ufności każdej dokonanej przez algorytm klasyfikacji. W ramach eksperymentów dokonano szczegółowej analizy wyników, korzystając z takich metryk jak dokładność detekcji (tj. procent poprawnych klasyfikacji ogółem), czułość (procent poprawnie wykrytych materiałów zmanipulowanych) oraz specyficzność (procent poprawnie zakwalifikowanych materiałów autentycznych).

Szczególna uwaga została poświęcona analizie przypadków błędnych klasyfikacji, zarówno w zakresie fałszywych alarmów (*false positives*), czyli przypadków oznaczenia materiału autentycznego jako zmanipulowanego, jak i pominięć (*false negatives*), czyli sytuacji, w których algorytm nie wykrył faktycznie zmanipulowanego materiału. W ramach eksperymentów zbadano również efektywność poszczególnych metod w warunkach utrudnionych, takich jak pogorszona jakość nagrań spowodowana kompresją danych, która jest typowa dla materiałów rozpowszechnianych w mediach społecznościowych.

Eksperymenty przeprowadzone zgodnie z przedstawionymi założeniami metodologicznymi dostarczyły precyzyjnych i wiarygodnych wyników empirycznych. Pozwoliły one nie tylko ocenić skuteczność wybranych narzędzi detekcyjnych, ale również dostarczyły danych potrzebnych do sformułowania dalszych wniosków oraz rekomendacji. Uzyskane rezultaty potwierdziły, iż zaawansowane algorytmy bazujące na głębokich sieciach neuronowych wykazują bardzo wysoką skuteczność w wykrywaniu manipulacji audiowizualnych, jednak nadal ich efektywność zależy od wielu czynników zewnętrznych, takich jak jakość materiału źródłowego, sposób przeprowadzenia manipulacji, czy czas od publikacji do analizy. Eksperymenty te stanowią solidną podstawę naukową do dalszych analiz, jak również mają znaczenie praktyczne – mogą być bowiem podstawą do wdrażania tego typu technologii w instytucjach odpowiedzialnych za bezpieczeństwo informacyjne państwa. Wyniki tych badań, szczegółowo przedstawione w dalszej części pracy, jednoznacznie wskazują na konieczność kontynuowania prac nad udoskonalaniem technologii analitycznych, które będą w stanie skutecznie przeciwdziałać coraz bardziej wyrafinowanym formom manipulacji audiowizualnej.

7.3. Analiza wyników badań – porównanie wykrywalności manipulacji

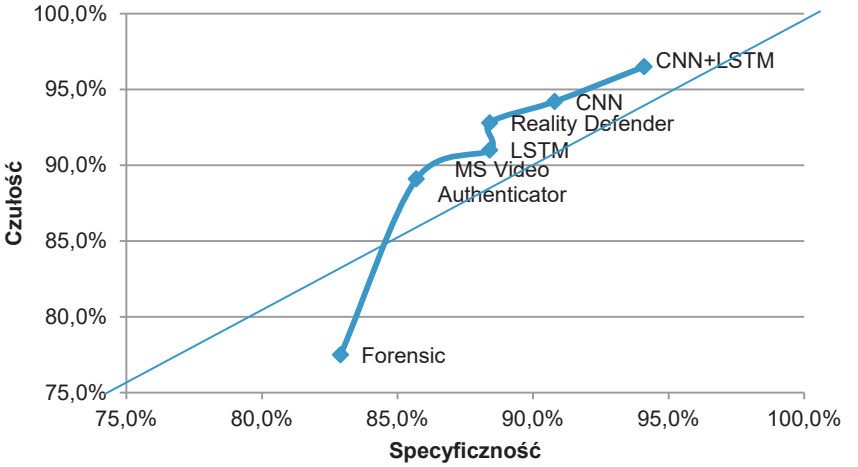
Przeprowadzone w ramach niniejszej pracy badania empiryczne pozwoliły na uzyskanie wyników dotyczących skuteczności różnych metod detekcji manipulacji audiowizualnych. Analiza tych wyników umożliwiła identyfikację mocnych oraz słabych stron poszczególnych algorytmów i metod detekcyjnych, dostarczając cennych informacji na temat ich użyteczności w kontekście bezpieczeństwa informacyjnego państwa. Szczegółowe zestawienie wyników badań zostało zaprezentowane w tabeli 6, stanowiącej punkt wyjścia do dalszej pogłębionej analizy.

Tabela 6. Wyniki eksperymentalnych testów skuteczności algorytmów detekcji manipulacji audiowizualnych

Metoda / algorytm	Dokładność	Czułość	Specyficzność	False Positive Rate (FPR)	False Negative Rate (FNR)
Konwolucyjna sieć neuronowa (CNN)	92,5%	94,2%	90,8%	9,2%	5,8%
Rekurencyjna sieć neuronowa (LSTM)	89,7%	91,0%	88,4%	11,6%	9,0%
Metoda forensyczna (analiza metadanych i artefaktów kompresji)	80,2%	77,5%	82,9%	17,1%	22,5%
Hybrydowy model CNN+LSTM	95,3%	96,5%	94,1%	5,9%	3,5%
Microsoft Video Authenticator	87,4%	89,1%	85,7%	14,3%	10,9%
Reality Defender	90,6%	92,8%	88,4%	11,6%	7,2%

Źródło: opracowanie własne.

Na rysunku 5 przedstawiono graficzne zestawienie wyników testów algorytmów, uwzględniające zależność między czułością a specyficznością poszczególnych metod detekcji manipulacji audiowizualnych. Linia diagonalna umieszczona na wykresie reprezentuje punkt równowagi pomiędzy tymi parametrami, umożliwiając łatwiejszą ocenę ogólnej efektywności poszczególnych metod.



Rysunek 5. Porównanie skuteczności badanych metod analizy audiowizualnej w detekcji manipulacji – zależność między czułością a specyficznością
Źródło: opracowanie własne.

Analizując szczegółowo wyniki przedstawione w tabeli 6 oraz na rysunku 5 można zaobserwować znaczne zróżnicowanie efektywności poszczególnych rozwiązań w zakresie detekcji manipulacji audiowizualnych. Zestawienie uzyskanych parametrów (dokładności, czułości, specyficzności oraz miar błędu klasyfikacyjnego) ukazuje wyraźną przewagę modelu multimodalnego (CNN+LSTM) nad pozostałymi metodami, co sugeruje, że równoczesna analiza zarówno warstwy wizualnej, jak i dźwiękowej materiałów wideo jest kluczowa w wychwytywaniu coraz bardziej złożonych i zaawansowanych technicznie fałszerstw. Przede wszystkim należy zwrócić uwagę na najwyższą wartość dokładności (95,3%) osiągniętą przez model hybrydowy CNN+LSTM. Osiągnięcie tak wysokiego poziomu poprawności klasyfikacji wskazuje, że w większości przypadków algorytm był w stanie prawidłowo sklasyfikować materiał jako zmanipulowany bądź autentyczny, co potwierdza jego przydatność w kontekście bezpieczeństwa informacyjnego. Dodatkowo wysoka czułość (96,5%) oznacza, że model ten wyjątkowo skutecznie wykrywał nagrania faktycznie fałszowane – czyli w niewielu sytuacjach fałszywe treści zostały zaklasyfikowane jako prawdziwe (ang. *false negatives*). Jednocześnie warto wskazać, że specyficzność na poziomie 94,1% również zapewnia względnie niewielki odsetek przypadków, w których nagranie autentyczne zostało błędnie uznane za zmanipulowane

(ang. *false positives*). Tak korzystne wyniki modelu CNN+LSTM korespondują z teoretycznymi podstawami metody multimodalnej. Konwolucyjne sieci neuronowe (CNN) bardzo dobrze radzą sobie z analizą statycznych atrybutów wizualnych, np. wykrywaniem nienaturalnych zniekształceń konturów, niespójności w oświetleniu czy artefaktów powstałych podczas edycji materiału wideo. Z kolei sieci rekurencyjne LSTM (ang. *Long Short-Term Memory*) pozwalają na uwzględnienie sekwencyjnego charakteru danych audio, co ułatwia identyfikację anomalii w modulacji głosu oraz dynamicznych zmian dźwięku. Połączenie obu tych podejść – analizującego jednocześnie obraz i dźwięk – powoduje efekt synergii, który w rezultacie przekłada się na mniejszą liczbę błędnych klasyfikacji zarówno w kategoriach *false positives*, jak i *false negatives*. Modele wykorzystujące pojedynczą modalność (CNN oraz LSTM analizowane osobno) wypadają zauważalnie słabiej w zakresie ogólnej dokładności, plasując się odpowiednio na poziomie 92,5% (CNN) i 89,7% (LSTM). Różnica względem rozwiązań multimodalnych (ok. 3-5 punktów procentowych) pozostaje istotna z punktu widzenia praktycznego wdrożenia, ponieważ w przypadku systemów bezpieczeństwa nawet kilka dodatkowych procent może oznaczać wyraźne podniesienie skuteczności wykrywania potencjalnie niebezpiecznych incydentów.

Zdecydowanie niższe wyniki odnotowały klasyczne metody cyfrowej analizy forensycznej (80,2% dokładności), dla których obserwowany poziom czułości (77,5%) i specyficzności (82,9%) jest znacząco niższy niż przy podejściach bazujących na głębokich sieciach neuronowych. Techniki forensyczne tradycyjnie koncentrują się na wyszukiwaniu artefaktów wynikających z procesu edycji lub konwersji plików (np. analiza metadanych, wykrywanie charakterystycznych wzorców w kompresji), jednak coraz bardziej wyrafinowane metody generowania materiałów typu *deepfake* potrafią tworzyć wiarygodne zniekształcenia, które nie zostawiają w metadanych oczywistych „śladów" edycji. Dlatego właśnie ich wskaźnik *false positives* (17,1%) oraz *false negatives* (22,5%) jest wysoki, co w praktyce oznacza częste błędne klasyfikacje zarówno w kierunku mylnego oskarżania autentycznych materiałów o sfałszowanie, jak i przeoczenia faktycznie zmanipulowanych przekazów.

Warto również zwrócić uwagę na wielkość wskaźnika błędnych klasyfikacji w metodach bazujących wyłącznie na uczeniu głębokim (CNN lub LSTM). Choć ich wartości FPR (9,2% i 11,6%) oraz FNR (5,8% i 7,2%) są ogólnie niższe niż w przypadku rozwiązań forensycznych, to nadal mogą

generować poważne konsekwencje praktyczne w kontekście służb odpowiedzialnych za bezpieczeństwo państwa czy instytucji *fact-checkingowych.* W sytuacjach krytycznych, np. podczas szybkiej weryfikacji nagrań stanowiących potencjalne źródło zamieszek czy paniki, każda pomyłka klasyfikacyjna może powodować realne skutki społeczne lub polityczne.

Podsumowując, przedstawione dane liczbowe w sposób jednoznaczny wskazują na wyraźną przewagę rozwiązań multimodalnych nad podejściami opartymi na pojedynczej modalności lub klasycznych technikach forensycznych. Model hybrydowy CNN+LSTM demonstruje najwyższą efektywność detekcji manipulacji audiowizualnych i jednocześnie najniższe wskaźniki błędów klasyfikacyjnych, co czyni go w chwili obecnej najlepszą opcją dla wdrożeń w strukturach bezpieczeństwa informacyjnego. Niemniej jednak, nawet rozwiązania osiągające ok. 95% dokładności wciąż generują pewien margines błędu, który w skrajnych sytuacjach mógłby być brzemienny w skutkach. Stąd postuluje się dalsze prace nad optymalizacją istniejących algorytmów, a także nad tworzeniem jeszcze bardziej zaawansowanych modeli zdolnych do wykrywania przyszłych, bardziej złożonych rodzajów manipulacji, jakie z pewnością będą się pojawiać wraz z rozwojem technologii *deepfake* i pokrewnych. Wyniki tych badań potwierdzają zatem, że w kontekście bezpieczeństwa państwa, szeroko pojęte metody sztucznej inteligencji – zwłaszcza w wydaniu hybrydowym, wykorzystującym zarówno analizę obrazu, jak i dźwięku – stanowią perspektywiczną i efektywną odpowiedź na wyzwania związane z narastającą falą dezinformacji audiowizualnej. W kolejnych rozdziałach przedstawiono rekomendacje dotyczące praktycznego zastosowania tych algorytmów w systemach bezpieczeństwa, a także propozycje rozwoju i wzmocnienia instrumentów legislacyjnych, które w połączeniu z nowoczesną technologią mogą istotnie podnieść odporność państwa na manipulacje w obszarze informacyjnym.

7.4. Wpływ zastosowanych metod na zwiększenie odporności państwa na dezinformację

Analiza przeprowadzonych badań empirycznych oraz szczegółowe porównanie skuteczności poszczególnych metod wykrywania manipulacji audiowizualnych pozwalają na stwierdzenie, że zastosowanie nowoczesnych

technologii analitycznych odgrywa istotną rolę w budowaniu odporności państwa na zagrożenia informacyjne. W szczególności wykorzystanie metod multimodalnych, opartych na zaawżowanych algorytmach sztucznej inteligencji (takich jak konwolucyjne oraz rekurencyjne sieci neuronowe – CNN i LSTM), może znacząco wpłynąć na podniesienie skuteczności systemów bezpieczeństwa informacyjnego, przyczyniając się do ograniczenia ryzyka i skutków kampanii dezinformacyjnych.

W kontekście polskim problematyka ta nabiera szczególnego znaczenia z uwagi na intensyfikację działań manipulacyjnych skierowanych wobec instytucji publicznych oraz polskiego społeczeństwa. Rzeczywistość współczesnej przestrzeni medialnej, zdominowanej przez szybkie rozpowszechnianie materiałów audiowizualnych w Internecie, stanowi poważne wyzwanie dla krajowych służb odpowiedzialnych za bezpieczeństwo państwa. W Polsce aktualnie istniejące struktury instytucjonalne – takie jak Agencja Bezpieczeństwa Wewnętrznego, Agencja Wywiadu czy Służba Kontrwywiadu Wojskowego – powinny uwzględnić w swoich strukturach dedykowane komórki zajmujące się analizą i przeciwdziałaniem zagrożeniom informacyjnym. Zastosowanie zaawansowanych narzędzi detekcyjnych, takich jak opisane w badaniach modele CNN oraz LSTM, pozwala na zwiększenie precyzji w rozpoznawaniu fałszerstw audiowizualnych, a tym samym na szybsze przeciwdziałanie próbom manipulacji opinią publiczną.

Aby w sposób efektywny zastosować przedstawione rozwiązania z zakresu wykrywania i przeciwdziałania audiowizualnym manipulacjom, polskie instytucje państwowe odpowiedzialne za bezpieczeństwo powinny podjąć kompleksowe, wielostopniowe działania o charakterze wdrożeniowym. Jednym z kluczowych kroków jest rozbudowa wyspecjalizowanych struktur w służbach specjalnych, takich jak Agencja Bezpieczeństwa Wewnętrznego (ABW), Agencja Wywiadu (AW) czy Służba Kontrwywiadu Wojskowego (SKW). Konieczne staje się utworzenie nowych komórek analitycznych ukierunkowanych wyłącznie na bezpieczeństwo informacyjne i wyposażenie ich w zaawansowane oprogramowanie przeznaczone do analizy materiałów medialnych. Przykładem mogą być systemy oparte na sieciach konwolucyjnych oraz innych algorytmach sztucznej inteligencji, zdolne do monitorowania przestrzeni internetowej pod kątem *deepfake'ów* oraz pozostałych form manipulacji wizualnej i dźwiękowej. Zaleca się również stworzenie scentralizowanego ośrodka analitycznego w zakresie dezinformacji

(na przykład w strukturach Agencji Bezpieczeństwa Wewnętrznego lub w obrębie Rządowego Centrum Bezpieczeństwa), który koordynowałby zgłoszenia o podejrzanych nagraniach z różnych źródeł – zarówno służb, ministerstw, mediów, jak i samych obywateli. Dzięki wykorzystaniu profesjonalnych narzędzi analitycznych ośrodek ten mógłby pracować w trybie całodobowym, pełniąc funkcję systemu wczesnego ostrzegania. W sytuacji zidentyfikowania istotnego zagrożenia, takiego jak zorganizowana kampania dezinformacyjna, miałby on możliwość natychmiastowego ostrzeżenia odpowiednich organów państwowych oraz przygotowania odpowiednich działań o charakterze kontr-narracyjnym (np. wydania oficjalnego sprostowania lub zorganizowania konferencji prasowej, której celem byłoby zaprezentowanie dowodów obalających fałszywkę).

Istotną rolę w tym procesie pełni również integracja rozwiązań technologicznych w istniejących już strukturach związanych z cyberbezpieczeństwem. W praktyce narzędzia detekcyjne muszą zostać połączone z mechanizmami, którymi dysponuje *Computer Emergency Response Team* Polska (CERT Polska) oraz z systemami monitoringu mediów społecznościowych działającymi na przykład w Ministerstwie Cyfryzacji i w Ministerstwie Spraw Zagranicznych. Celem jest usprawnienie przepływu informacji o incydentach w sferze informacyjnej i potraktowanie ich z podobnym priorytetem, jaki dotychczas zarezerwowany był głównie dla ataków *hakerskich*. W dalszej kolejności szczególnie ważne staje się inwestowanie w edukację kadr – analitycy odpowiedzialni za identyfikację i neutralizację materiałów manipulacyjnych powinni dysponować aktualną wiedzą na temat najnowszych metod fałszowania (np. zaawansowanego klonowania głosu i generowania obrazu) oraz potrafić płynnie wykorzystywać dostępne narzędzia wykrywające te zagrożenia. Kluczowe będzie tu nawiązanie współpracy z ośrodkami naukowymi i przedsiębiorstwami oferującymi nowoczesne rozwiązania technologiczne; podkreślić należy zwłaszcza znaczenie partnerstw badawczo-rozwojowych, między innymi z polskimi projektami tego typu (przykładowo z inicjatywą „IDEAS" Narodowego Centrum Badań i Rozwoju), które już stworzyły prototypowe systemy służące do wykrywania głębokich fałszerstw (*deepfake*). Takie partnerstwo daje państwowym instytucjom dostęp do bieżących osiągnięć naukowych opracowywanych w kraju, jednocześnie przyspieszając proces wdrożeń w służbach i organach administracji publicznej.

Równoległe działania powinny zostać podjęte w obszarze mediów publicznych, które – ze względu na duży zasięg oraz szczególny mandat w zakresie edukowania i informowania obywateli – mają znaczący wpływ na kształtowanie odporności społeczeństwa na dezinformację. Zaleca się, aby redakcje Telewizji Polskiej i Polskiego Radia wprowadziły do swoich procedur zaawansowane narzędzia forensyczne, zdolne do precyzyjnego sprawdzania autentyczności materiałów przed ich opublikowaniem. Rozwiązania te, w formie różnego rodzaju wtyczek (plug-inów) i specjalistycznych programów, pozwalają rozpoznać typowe artefakty pozostawiane przez systemy generujące *deepfake*. W dalszej perspektywie media publiczne powinny wzmocnić swój potencjał ekspercki przez powołanie zespołów weryfikatorów, których zadaniem byłaby stała ocena wiarygodności wszelkich dostarczanych treści (z uwzględnieniem możliwych ingerencji w głos, obraz czy montaż nagrań). W sytuacjach kryzysowych, kiedy rozpowszechniane są nieprawdziwe materiały o znaczeniu strategicznym, media publiczne mogłyby w trybie przyspieszonym oferować rzetelną analizę i prostować fałszywe informacje – stanowiąc w ten sposób ważny filar oficjalnej, opartej na dowodach i przejrzystej komunikacji. Równolegle warto, aby media państwowe podejmowały współpracę z niezależnymi platformami typu *fact-checking*, na przykład w formie wspólnych akcji uświadamiających czy kampanii edukacyjnych skierowanych do szerokiego grona odbiorców. Dodatkowo, organy nadzorujące rynek medialny, takie jak Krajowa Rada Radiofonii i Telewizji, mogłyby opracować lub wzmocnić wytyczne, zobowiązujące nadawców do precyzyjnego oznaczania syntetycznie wygenerowanych materiałów (np. opatrzenia stosowną informacją w rogu ekranu bądź w podpisie), co wpłynie na zwiększenie przejrzystości przekazu i ograniczy ryzyko niezamierzonego wprowadzania widzów czy słuchaczy w błąd.

Jednocześnie w ministerstwach i innych jednostkach administracji publicznej należy wdrożyć jasne procedury reagowania na zidentyfikowane przypadki dezinformacji audiowizualnej. Każda kluczowa instytucja, taka jak Ministerstwo Obrony Narodowej, Ministerstwo Spraw Zagranicznych czy Ministerstwo Spraw Wewnętrznych i Administracji, powinna stworzyć wewnętrzne zespoły do spraw komunikacji kryzysowej, które w razie wykrycia fałszywego materiału odnoszącego się do kompetencji danego resortu (przykładowo spreparowanej wypowiedzi Ministra Obrony Narodowej) podjęłyby natychmiastową weryfikację autentyczności nagrania

we współpracy ze służbami wyspecjalizowanymi. Następnym krokiem byłoby niezwłoczne wydanie oficjalnego oświadczenia, w którym stanowczo zaprzecza się rozpowszechnianej dezinformacji i prezentuje wnioski płynące z przeprowadzonej analizy. Na poziomie systemowym potrzebne są też szczegółowe protokoły postępowania na wypadek wystąpienia głębokich fałszerstw – ujednolicą one i skoordynują działania rządu, zapewniając spójną linię komunikacyjną (np. jednoczesne ogłoszenie dementi w mediach społecznościowych, na rządowych portalach informacyjnych i w ramach konferencji prasowej). Istotne jest także konsekwentne prowadzenie monitoringu przestrzeni informacyjnej w obszarach kluczowych dla bezpieczeństwa państwa. Przykładowo Ministerstwo Zdrowia, działające w czasie kryzysów zdrowotnych, powinno śledzić publikowane w sieci materiały dotyczące epidemiologii i ochrony zdrowia, aby w porę rozpoznać i zdementować ewentualne wypowiedzi rzekomych ekspertów, które mogą wywoływać nieuzasadnioną panikę. Podobne działania powinno realizować Ministerstwo Obrony Narodowej w kontekście doniesień dotyczących granic lub przedsięwzięć wojskowych. W tym zakresie kluczową rolę odgrywa wymiana informacji z centralnym ośrodkiem analitycznym, a także możliwość wdrożenia autorskich systemów automatycznego filtrowania treści według słów kluczowych i analizy za pomocą algorytmów sztucznej inteligencji. Wzmocnieniu wiarygodności przekazu służą oficjalne kanały komunikacji rządu (strony internetowe, profile w mediach społecznościowych), które – jeśli są starannie prowadzone i transparentnie prezentują dane – mogą skutecznie przeciwdziałać rozchodzeniu się nieprawdziwych informacji. Regularne raporty na temat skutecznie zidentyfikowanych i zneutralizowanych kampanii dezinformacyjnych wzmacniają społeczne zaufanie i pokazują, że państwo aktywnie dba o bezpieczeństwo informacyjne obywateli.

Problematyka wykrywania i zwalczania zaawansowanej dezinformacji posiada wyraźny kontekst międzynarodowy, co wymaga szerokiej i wielostronnej współpracy. Polska powinna zatem kontynuować oraz intensyfikować współdziałanie z innymi państwami, jak również z organizacjami międzynarodowymi. Na forum Unii Europejskiej temat dezinformacji jest uznawany za istotny element zagrożeń hybrydowych, co skutkuje rozmaitymi inicjatywami. Działania w tym obszarze prowadzi chociażby unijna grupa zadaniowa *East StratCom*, funkcjonująca przy Europejskiej Służbie Działań Zewnętrznych od 2015 roku, której celem jest śledzenie i ujawnianie

propagandy prokremlowskiej na obszarze Europy. Ponadto wspólne ramy prawne i koncepcyjne w zakresie reagowania na zagrożenia hybrydowe, tworzone pod auspicjami Unii Europejskiej, ułatwiają podejmowanie skoordynowanych działań przez państwa członkowskie. Polska, jako członek Unii Europejskiej, bierze udział w systemie wymiany informacji wywiadowczych dotyczących wykrytych kampanii manipulacyjnych, między innymi poprzez System Wczesnego Ostrzegania UE ds. dezinformacji, ustanowiony w 2019 roku. Równocześnie, dzięki udziałowi w opracowywaniu unijnych regulacji (np. Kodeksu postępowania w sprawie dezinformacji), Polska może domagać się od międzynarodowych platform internetowych (w tym takich firm, jak Meta, Google czy X – dawniej Twitter) usprawnienia procesów zgłaszania i usuwania treści nieprawdziwych, a także objęcia zaawansowanych technik manipulacji (w tym *deepfake*) szczególnym nadzorem. Wzmacniają to sukcesywnie wprowadzane akty prawne, między innymi Akt o Usługach Cyfrowych (Digital Services Act) czy projektowane przepisy Aktu o Sztucznej Inteligencji, które nakładają na platformy obowiązek informowania użytkowników o faktach generowania treści w sposób syntetyczny.

Nie mniej znacząca jest współpraca w ramach Organizacji Traktatu Północnoatlantyckiego (NATO), gdzie dezinformację postrzega się jako jeden z elementów wojny hybrydowej. Przejawem tego jest funkcjonowanie Centrum Doskonalenia Komunikacji Strategicznej NATO (tzw. *StratCom CoE* w Rydze), analizującego między innymi wpływ sztucznej inteligencji na propagandę i zagrożenia nią wywoływane. Polska jako kraj sojuszniczy, uczestniczy w wymianie doświadczeń i informacji wywiadowczych, dzięki czemu może wykorzystywać sprawdzone rozwiązania wypracowane przez partnerów zagranicznych – szczególnie tych, którzy mierzą się z agresywną dezinformacją od dłuższego czasu (przykład stanowią państwa bałtyckie). Dodatkowym przykładem tej międzynarodowej kooperacji jest polsko-amerykańska inicjatywa o nazwie *Ukraine Communications Group*, uruchomiona w 2024 roku w Warszawie i skupiająca przedstawicieli dwunastu państw zachodnich, w tym Stanów Zjednoczonych, których celem jest wspólne wykrywanie oraz demaskowanie manipulacji wymierzonych w Ukrainę. Grupa ta, wspierana przez Departament Stanu USA, opracowuje narzędzia monitorujące narracje szerzone w mediach i ustala strategie zwalczania szkodliwych kampanii informacyjnych. Polski udział w takich projektach stanowi szansę na korzystanie z bogatych zasobów danych i najnowszych

osiągnięć technologicznych, a także na usprawnianie własnych metod identyfikacji i obalania fałszywych materiałów.

W dalszej perspektywie polska dyplomacja może też współuczestniczyć w ustalaniu standardów prawnych na szczeblu globalnym. Coraz częściej pojawiają się postulaty, aby w ramach Organizacji Narodów Zjednoczonych (ONZ) i innych organizacji międzynarodowych wprowadzić przepisy i normy potępiające wykorzystywanie zmanipulowanych materiałów audiowizualnych do destabilizowania sytuacji politycznej w innych państwach. Choć zbudowanie międzynarodowego konsensusu w tej dziedzinie stanowi długofalowy proces, już samo podnoszenie zagadnienia w debacie globalnej przyczynia się do wzrostu świadomości i sprzyja wypracowywaniu stopniowych rozwiązań. Pozytywne rezultaty mogą przynieść także działania polegające na nawiązywaniu partnerstw z sektorem pozarządowym i prywatnym, zwłaszcza z koncernami technologicznymi działającymi na skalę światową. Firmy te coraz częściej deklarują gotowość do wzajemnego dzielenia się informacjami o nowopowstałych technikach generowania lub wykrywania *deepfake'ów,* aby wspólnie ograniczać negatywne skutki szerzenia się treści zmanipulowanych. Z kolei niezależne sieci *fact-checkerów* (takie jak *International Fact-Checking Network*) regularnie ostrzegają się nawzajem o trendach w dezinformacji zaobserwowanych w różnych regionach świata. Zaangażowanie Polski w te globalne inicjatywy dodatkowo wzmacnia odporność państwa na zagrożenia informacyjne, pozwalając czerpać z doświadczeń międzynarodowych i korzystać z szerokiej bazy zasobów eksperckich.

Podsumowując, wzajemne przenikanie się czynników organizacyjnych, technologicznych i edukacyjnych w zakresie przeciwdziałania audiowizualnym manipulacjom wymaga od Polski kompleksowego i długofalowego podejścia. Kluczowe jest stworzenie rozbudowanych struktur analitycznych w służbach specjalnych, uspójnienie narzędzi cyberbezpieczeństwa i zwiększenie kompetencji pracowników zajmujących się bezpieczeństwem informacyjnym, a także pełna integracja mediów publicznych z krajowym ekosystemem wykrywania manipulacji. Konieczna pozostaje również ścisła współpraca międzynarodowa na poziomie Unii Europejskiej, Organizacji Traktatu Północnoatlantyckiego oraz w ramach kontaktów dwustronnych i wielostronnych, co pozwala na efektywną wymianę wiedzy, doświadczeń i technologii. Tylko przy takim wielokierunkowym działaniu można oczekiwać, że państwo będzie w stanie przeciwdziałać nawet

najbardziej wyrafinowanym kampaniom dezinformacyjnym, chroniąc tym samym społeczeństwo przed konsekwencjami fałszowania oraz nadużywania treści o charakterze audiowizualnym.

VIII

Wyzwania i przyszłość technologii wykrywania dezinformacji

8.1. Ograniczenia technologiczne w analizie treści audiowizualnych

Analiza treści audiowizualnych z wykorzystaniem sztucznej inteligencji stała się kluczowym elementem współczesnych systemów bezpieczeństwa. Mimo imponujących postępów w dziedzinie uczenia głębokiego, istnieje szereg ograniczeń technologicznych, które utrudniają w pełni niezawodne zastosowanie tych rozwiązań w praktyce. Pierwszym wyzwaniem jest skala danych oraz wymóg przetwarzania ich w czasie rzeczywistym. W erze big data systemy monitoringu generują ogromne ilości nagrań – od miejskich kamer CCTV po strumienie wideo z mediów społecznościowych. Już kilkanaście lat temu zauważono, że rozmiary systemów dozoru wizyjnego przekraczają możliwości manualnej analizy, a automatyzacja napotyka problemy zapewnienia wysokiej wykrywalności przy jednoczesnym niskim odsetku fałszywych alarmów[285]. Współczesne algorytmy *video analytics* muszą przetwarzać setki klatek na sekundę, co wymaga znacznej mocy obliczeniowej i wydajnych infrastrukturalnie rozwiązań sprzętowych. Zapewnienie działania w trybie rzeczywistym bywa trudne – opóźnienia wynikające z tzw. bariery inferencji (tj. ograniczeń w szybkim wnioskowaniu modeli AI na nowych danych) mogą uniemożliwiać natychmiastową reakcję na incydenty. Symulacje pokazują, że system analizujący obraz z kilkuset kamer jednocześnie generuje lawinowo zdarzenia do weryfikacji, nawet przy niewielkim

[285] M. Valera, S.A. Velastin, Intelligent Distributed Surveillance Systems: a review, IEE Proceedings – Vision, Image & Signal Processing, 152(2), 192-204.

odsetku błędnych detekcji. Przykładowo, przy 0,5% fałszywych alarmów, na każde 1000 zdarzeń oznaczonych przez algorytm jako podejrzane aż 5 będzie fałszywymi sygnałami – co przy tysiącach analizowanych klipów dziennie daje dziesiątki fałszywych alarmów. Zalew takiej liczby powiadomień przekracza zdolności reagowania operatorów, prowadząc do tzw. zmęczenia alarmami i ryzyka ignorowania nawet prawdziwych zagrożeń. Problem skali dotyczy również samego przesyłu i przechowywania danych. Nagrania w wysokiej rozdzielczości zajmują olbrzymią przestrzeń dyskową, a ich ciągłe transmitowanie do centralnych serwerów analitycznych wymaga ogromnej przepustowości sieci. W praktyce wymusza to stosowanie kompresji (co może obniżać jakość obrazu) lub przeniesienie części obliczeń bliżej źródła danych (analiza na brzegu sieci, tzw. *edge computing*). Kamery wyposażone we wbudowane moduły AI mogą wstępnie filtrować dane, jednak takie rozwiązania również mają ograniczenia – muszą działać na energooszczędnych procesorach o małej mocy, co ogranicza złożoność modeli, jakie da się na nich uruchomić. Ponadto, zdecentralizowana analiza utrudnia globalną korelację informacji z wielu źródeł jednocześnie. Kolejnym ograniczeniem jest dokładność i niezawodność automatycznej interpretacji obrazu i dźwięku. Algorytmy uczenia maszynowego, choć potrafią wykrywać obiekty i zdarzenia z wysoką precyzją na danych testowych, w warunkach rzeczywistych często doświadczają spadku skuteczności. Nawet niewielkie zmiany warunków oświetlenia, pogody, kąta kamery czy jakości nagrania mogą spowodować błędne rozpoznania.

W kontekście bezpieczeństwa oznacza to ryzyko zarówno fałszywych alarmów – sygnalizowanie zagrożenia tam, gdzie go nie ma, jak i przeoczeń faktycznego niebezpieczeństwa. Przykładowo, subtelne różnice w zachowaniu tłumu – które dla człowieka są oczywiste w interpretacji jak np. panika vs. zwykły pośpiech – dla systemu mogą być trudne do odróżnienia. Problem ten wiąże się z tzw. „luką semantyczną", czyli różnicą między niskopoziomowymi cechami obrazu wychwytywanymi automatycznie a ich wysokopoziomowym znaczeniem, które nadaje im człowiek[286]. Innymi słowy, system potrafi zidentyfikować, że w kadrze znajduje się biegnąca postać, lecz nie rozumie kontekstu – czy osoba biegnie w celach rekreacyjnych, czy ucieka

[286] Zob. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349-1380.

z miejsca przestępstwa. W praktyce zdarzały się przypadki, że algorytm oceniał niegroźne sytuacje jako zagrożenie lub odwrotnie, np. intensywną dyskusję omyłkowo interpretował jako akt agresji. Niejednoznaczność kontekstu pozostaje barierą, której obecna technologia nie pokonuje w satysfakcjonującym stopniu. Ograniczenia dotyczą także samych modeli sztucznej inteligencji. Nowoczesne systemy rozpoznawania obrazu jak detekcji twarzy czy identyfikacji podejrzanych zachowań opierają się na złożonych sieciach neuronowych, które działają jak „czarna skrzynka". Brak przejrzystości co do sposobu podjęcia decyzji przez algorytm rodzi wyzwania z punktu widzenia zaufania i odpowiedzialności. W systemach bezpieczeństwa jest to szczególnie istotne – trudno zaakceptować wynik klasyfikacji jako podstawę działania (np. zatrzymania podejrzanego) bez możliwości wyjaśnienia, dlaczego algorytm uznał go za zagrożenie. W odpowiedzi na te obawy pojawia się postulat *Explainable AI*, lecz implementacja takich rozwiązań w domenie analizy wideo wciąż jest we wczesnym etapie rozwoju[287]. Co więcej, same modele są wrażliwe na błędne dopasowanie do danych treningowych – mogą nauczyć się artefaktów specyficznych dla nagrań z jednego miasta czy jednego typu kamery, przez co przeniesione w nowe środowisko tracą skuteczność. Każda zmiana scenerii, inny układ ulic, inna kultura zachowań społecznych, może wymagać czasochłonnego dostrajania modeli lub ponownego ich trenowania. Następnym wyzwaniem jest jakość i reprezentatywność danych treningowych, na których uczone są modele. Algorytmy uczą się na podstawie dostarczonych im przykładów – jeśli w danych występują uprzedzenia lub luki, system je przejmie. W kontekście analizy audiowizualnej może to prowadzić do stronniczości algorytmów. Głośnym przykładem jest niższa skuteczność komercyjnych systemów rozpoznawania twarzy wobec kobiet oraz mniejszości etnicznych, co wykazano w badaniach nad dysproporcjami w dokładności klasyfikacji[288]. Tego typu obciążenia mogą skutkować nieproporcjonalnym wskazywaniem niektórych grup jako podejrzanych lub generować błędne identyfikacje, rodząc poważne implikacje etyczne i prawne. Jeśli system nadzoru częściej błędnie identyfikuje osoby o określonym wyglądzie jako poszukiwanych przestępców, może to prowadzić do ich niesłusznego zatrzymania

[287] High-Level Expert Group on AI (2019), Ethics Guidelines for Trustworthy AI, Komisja Europejska, Bruksela 2019.

[288] Zob. J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of Machine Learning Research, 81, 1-15.

lub nękania, co jest niedopuszczalne. Z kolei pewne zagrożenia bezpieczeństwa są na tyle rzadkie np. ataki terrorystyczne na lotniskach, że brak dużych zbiorów danych prezentujących takie sytuacje – modele nie mają się z czego uczyć lub opierają się na symulacjach, które nie oddają w pełni złożoności realnych zdarzeń. W efekcie skuteczność algorytmu w wykrywaniu bardzo nietypowych incydentów pozostaje niewiadomą aż do chwili, gdy rzeczywiście do nich dojdzie, co obniża zaufanie do systemu. Specyficznym, szybko rozwijającym się obszarem jest automatyczne wykrywanie manipulacji medialnych, takich jak głęboko fałszowane nagrania wideo i audio (ang. *deepfake*). Tutaj również ujawniają się ograniczenia technologiczne. Najnowocześniejsze metody detekcji *deepfake* opierają się na głębokich sieciach neuronowych trenowanych na dużych zbiorach przykładowych fałszerstw. Jednak badania wskazują, że obecne detektory mają ograniczoną skuteczność uogólniania – model wytrenowany na jednym zestawie sfałszowanych nagrań często nie radzi sobie z innymi, nowymi metodami fałszowania, które nie były reprezentowane w danych uczących. Jak zauważają Kaur i współautorzy, wyzwaniem są zarówno problemy z danymi (m.in. niezrównoważone lub szczątkowe zbiory treningowe), jak i aspekty obliczeniowe (konieczność ogromnej mocy obliczeniowej do trenowania i działania modeli detekcji) oraz ograniczona odporność na nowe ataki (pojawianie się coraz doskonalszych technik generowania fałszywych nagrań) – wszystko to sprawia, że automatyczne wykrywanie *deepfake* jest dalekie od doskonałości[289]. Ponadto, wykrywanie fałszerstw staje się trudniejsze wraz z poprawą jakości generowanych *deepfake'ów*. Techniki fałszowania ewoluują, a generowane multimedia coraz lepiej imitują rzeczywistość, przez co wyprzedzają zdolności detekcyjne dostępnych narzędzi. Jest to sytuacja swoistego „wyścigu zbrojeń" między twórcami manipulacji a ich wykrywaczami, w którym każde udoskonalenie detektorów spotyka się z kontrdziałaniem polegającym na tworzeniu trudniejszych do wykrycia fałszerstw. Przykładowo, pierwsze algorytmy wykrywające *deepfake* zwracały uwagę na nienaturalnie rzadkie mruganie powiek u osób na podrobionych nagraniach (co było ubocznym efektem ograniczeń zestawów treningowych), jednak nowsze generatory nauczyły się symulować prawidłowe mruganie, niwecząc tę metodę detekcji (Li i in., 2018). Podobnie, stosowane niegdyś proste detektory artefaktów

[289] Zob. A. Kaur, A.N. Hoshyar, V. Saikrishna, S. Firmin, F. Xia, Deepfake video detection: challenges and opportunities, Artificial Intelligence Review, 57, 159, 2024.

kompresji czy zniekształceń na krawędziach obiektów stały się nieskuteczne, gdyż nowe modele generatywne potrafią te artefakty eliminować. Oznacza to, że każda metoda wykrywania manipulacji ma ograniczony "okres przydatności" dopóty, dopóki nie zostanie zneutralizowana przez nową generację narzędzi fałszujących. Równie ważnym komponentem analizy audiowizualnej jest dźwięk – np. automatyczna transkrypcja mowy lub rozpoznawanie emocji w głosie. I na tym polu technologia ma swoje ograniczenia. W warunkach laboratoryjnych systemy rozpoznawania mowy osiągają już bardzo niskie współczynniki błędu – najlepsze modele potrafią uzyskać poniżej 5% błędnie rozpoznanych słów dla czystego nagrania[290]. Jednak w praktyce nagrania audio rzadko są idealnie czyste. Występuje szum tła, mówcy mają różne akcenty, dykcję, albo wtrącają słowa z innych języków, może też nakładać się mowa kilku osób jednocześnie. To wszystko powoduje drastyczny wzrost wskaźnika błędów słownych i spadek jakości transkrypcji. W hałaśliwym otoczeniu, np. na dworcu kolejowym czy podczas zgromadzenia publicznego, automatyczne systemy rozpoznawania mowy mogą okazać się zawodne bez specjalnych technik filtracji hałasu czy wspomagania przez analizę wizji między innymi odczytywanie ruchu warg przez kamery, co również jest przedmiotem badań. Z kolei analiza treści audio pod kątem wykrywania emocji, stanu napięcia czy prawdomówności mówcy pozostaje w fazie eksperymentalnej – ludzkie głosy są bardzo zróżnicowane i silnie zależne od kontekstu wypowiedzi, co utrudnia stworzenie niezawodnych algorytmów w tym zakresie. Zaawansowane systemy mogą np. oceniać tempo mowy i modulację głosu jako wskazówki stresu, ale rozróżnienie czy podniesiony głos wynika ze zdenerwowania, czy np. z hałasu otoczenia, bywa poza zasięgiem obecnych modeli. Na osobne omówienie zasługuje kwestia odporności systemów analizy na celowe oszustwa i ataki antagonistyczne. Coraz więcej badań dowodzi, że modele AI można wprowadzić w błąd poprzez przygotowanie specjalnie spreparowanych danych wejściowych. Dotyczy to zarówno obrazów, jak i audio. Nieznaczne modyfikacje pikseli obrazu, niewidoczne gołym okiem dla człowieka, potrafią diametralnie zmienić wynik klasyfikacji sieci neuronowej – przykładowo sprawić, że system nie rozpozna twarzy lub sklasyfikuje obiekt jako zupełnie inny. W kontekście bezpieczeństwa państwa

[290] Zob. Y. Li, M.C. Chang, S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong 2018, s. 1-7.

oznacza to, że potencjalny intruz może zastosować kamuflaż przeciwko AI – np. założyć odzież ze specjalnym wzorem lub emitować zakłócający dźwięk – by zmylić system monitoringu. Literatura przedmiotu opisuje m.in. eksperymenty z okularami posiadającymi nadruk, który sprawiał, że system rozpoznawania twarzy mylnie identyfikował tożsamość osoby, pozwalając jej niejako „przejść incognito" przed kamerami. Inne badania pokazały, że przypięcie małej, pozornie niegroźnej naklejki z wzorem do ubrania może skutecznie zafałszować detekcję obiektu – osoba z taką naklejką stawała się dla systemu „niewidzialna" albo była klasyfikowana jako zupełnie co innego. Tego typu ataki antagonistyczne stanowią poważne wyzwanie, gdyż wymagają od projektantów systemów bezpieczeństwa nie tylko doskonalenia samej analizy treści, ale i zabezpieczenia jej przed manipulacjami ze strony osób trzecich. Walka z takimi zagrożeniami wymaga ciągłego testowania systemów AI pod kątem ich odporności i wdrażania dodatkowych mechanizmów weryfikujących integralność danych np. sprawdzanie cyfrowych podpisów nagrań lub wykrywanie podejrzanych wzorców zakłóceń, co jednak samo w sobie jest trudnym zadaniem badawczym.

Podsumowując, ograniczenia technologiczne sprawiają, że analiza treści audiowizualnych nie jest narzędziem niezawodnym ani całkowicie autonomicznym. Wysoka złożoność obliczeniowa, niepewność interpretacji, podatność na zakłócenia oraz ewoluujące zagrożenia powodują, że systemy te muszą być wdrażane z dużą dozą ostrożności. W praktyce oznacza to konieczność łączenia możliwości AI z nadzorem i weryfikacją ze strony człowieka. Dopiero taka symbioza pozwala zwiększyć skuteczność monitoringu i analiz, jednocześnie minimalizując ryzyko błędów. W literaturze podkreśla się potrzebę zachowania tzw. *human-in-the-loop*, czyli udziału operatorów ludzkich w procesie decyzyjnym systemów bezpieczeństwa opartych na AI. Ostateczne decyzje – zwłaszcza te o poważnych konsekwencjach – powinny podlegać zatwierdzeniu przez człowieka, który może uwzględnić kontekst i dodatkowe informacje spoza zakresu „wiedzy" algorytmu. Ponadto wskazane jest kontynuowanie badań nad bardziej odpornymi, wydajnymi i wyjaśnialnymi algorytmami, a także gromadzenie lepszych jakościowo danych treningowych. Tylko w ten sposób ograniczenia opisane powyżej będą mogły być stopniowo przezwyciężane, pozwalając systemom analizy audiowizualnej w pełni wspierać bezpieczeństwo państwa.

8.2. Etyczne aspekty automatycznego wykrywania manipulacji medialnych

Wprowadzanie do systemów bezpieczeństwa państwa narzędzi automatycznie wykrywających manipulacje w przekazach medialnych – takich jak sfałszowane obrazy, nagrania wideo i audio – rodzi szereg dylematów etycznych. Technologie te obiecują zwiększyć odporność społeczeństwa na dezinformację i chronić obywateli przed szkodami wyrządzanymi przez fałszywe treści między innymi przed oszustwami z użyciem *deepfake* czy przed zniesławieniem z wykorzystaniem przerobionych nagrań. Skala problemu jest niebagatelna. Automatyczne wykrywanie takich treści i ich eliminowanie z platform mogłoby chronić godność i prywatność ofiar. Jednak zastosowanie nowych narzędzi niesie ze sobą ryzyko naruszenia podstawowych praw i wolności, na których opiera się demokratyczne państwo prawa. Pojawia się napięcie między bezpieczeństwem informacyjnym a prawem do prywatności, wolnością słowa oraz zasadami rzetelnego procesu.

Poniżej omówiono kluczowe wyzwania etyczne związane z automatyzacją wykrywania manipulacji medialnych. Pierwszym z nich jest prywatność i nadzór. Automatyczne skanowanie ogromnych ilości treści audiowizualnych w poszukiwaniu manipulacji może prowadzić do naruszenia prywatności obywateli. Aby wychwycić fałszywe nagrania, systemy musiałyby potencjalnie analizować wszystkie dostępne materiały – od filmów publikowanych w mediach społecznościowych po przekazy telewizyjne i prywatne nagrania krążące w sieci. Takie masowe monitorowanie treści budzi obawy o stworzenie infrastruktury panoptykonu cyfrowego, w którym każda wypowiedź i wizerunek podlega automatycznej inspekcji. Nawet jeśli intencją jest jedynie filtracja treści nieautentycznych, to przesiewanie komunikacji w poszukiwaniu manipulacji trudno pogodzić z zasadą minimalizacji ingerencji w dane osobowe. Wrażliwe mogą być zwłaszcza komunikatory prywatne – czy państwo, chcąc wyłapać np. *deepfake* audio wykorzystywane do wyłudzeń, może etycznie uzasadnić skanowanie rozmów głosowych obywateli? Z drugiej strony argumentuje się, że skoro manipulacje medialne same w sobie naruszają prywatność i dobra osobiste ofiar, to w imię ochrony prywatności należy stosować narzędzia pozwalające takie nadużycia wykrywać i usuwać.

Etyczny dylemat polega tu na wyważeniu prawa jednostki do niezakłóconej komunikacji z prawem innej jednostki do obrony przed medialnym atakiem na jej osobę. W praktyce oznacza to konieczność bardzo ostrożnego definiowania uprawnień instytucji stosujących takie narzędzia – np. dopuszczenia monitoringu treści tylko przy zaistnieniu uzasadnionego podejrzenia przestępstwa i pod kontrolą sądową. Ważna jest także kwestia anonimizacji. Czy systemy wykrywające manipulacje mogą to robić bez gromadzenia nadmiarowych danych o niewinnych użytkownikach? Techniki takie jak lokalna detekcja na urządzeniach użytkowników lub algorytmy przetwarzające jedynie zaszyfrowane charakterystyki multimediów mogłyby ograniczyć ingerencję w prywatność, ale ich zastosowanie jest dopiero badane. Kolejna kwestia to wolność słowa i ryzyko cenzury. Automatyczne wykrywanie i usuwanie rzekomych manipulacji stawia pytanie: kto i na jakiej podstawie definiuje, co jest manipulacją? W systemach demokratycznych wolność wypowiedzi jest chroniona, w tym także prawo do satyry, parodii czy wyrażania kontrowersyjnych opinii. Istnieje ryzyko, że algorytmy będą nadmiernie gorliwe i błędnie zaklasyfikują legalne treści jako zagrożenie. Głośnym przypadkiem z ostatnich lat było rozpowszechnienie zmanipulowanego nagrania przemówienia spikerki Izby Reprezentantów USA Nancy Pelosi w 2019 r. – materiał spowolniono i obniżono w nim ton głosu, by sprawiać wrażenie, że polityk mówi niewyraźnie (choć nie był to *deepfake* generowany przez sieci neuronowe).

Wiele osób uznało go za autentyczny, a platformy internetowe debatowały, czy taki „*cheap fake*" należy traktować jak zakazaną manipulację. Część serwisów jedynie oznaczyła wideo jako dezinformację, pozostawiając je online, argumentując prawem użytkowników do oceny satyry lub krytyki, podczas gdy inne zdecydowały się na jego usunięcie ze względu na szkodliwy, oszczerczy charakter przeróbki. Taka cenzura algorytmiczna byłaby trudna do zakwestionowania, zwłaszcza jeśli szczegóły działania detektora pozostają tajemnicą (brak transparentności algorytmu). To rodzi poważne obawy etyczne: czy w imię walki z dezinformacją nie ograniczymy nadmiernie swobody debaty publicznej? Szczególnie w państwach o słabszych instytucjach istnieje ryzyko nadużyć – władza może wykorzystywać narzędzia do wykrywania „fałszu" jako wygodne uzasadnienie tłumienia treści krytycznych wobec rządu. Już dziś w dyskursie publicznym pojawia się zjawisko

określane jako „dywidenda kłamcy”[291], gdzie osoby wpływowe podważają autentyczność prawdziwych materiałów dowodowych, nazywając je deepfake’ami, by uniknąć odpowiedzialności. Jeśli państwo dysponuje systemem rzekomo nieomylnie wykrywającym fałsz, pokusa może iść w drugą stronę – uznawać niewygodne, choć prawdziwe informacje za zmanipulowane. Dlatego tak ważne jest, by proces automatycznej weryfikacji był rzetelny, bezstronny i poddany niezależnej kontroli. Wolność słowa obejmuje także prawo do błędu czy nieścisłości – nie każde odstępstwo od prawdy jest złośliwą dezinformacją; automatyczne narzędzia muszą być więc stosowane z wyczuciem i w oparciu o jasne definicje prawne manipulacji medialnej. Rzetelność i transparentność algorytmów.

Etyka wymaga, aby systemy decydujące o losie czyjejś wypowiedzi działały w sposób sprawiedliwy i weryfikowalny. W przypadku automatycznego wykrywania manipulacji powstaje pytanie o wiarygodność dowodową wyniku takiej detekcji. Jeśli algorytm oceni, że dany film jest deepfake’em, czy twórca filmu ma prawo poznać uzasadnienie? Czy powinien istnieć tryb odwoławczy umożliwiający zakwestionowanie decyzji podjętej przez AI? W sferze moderacji treści internetowych coraz częściej postulowane jest wdrażanie mechanizmów „przejrzystego pudełka” zamiast „czarnej skrzynki” – tzn. ujawnianie kryteriów, którymi kieruje się algorytm. Brak transparentności może prowadzić do błędów i nadużyć pozostających w ukryciu. Z kolei pełna jawność metod detekcji stwarza ryzyko, że twórcy manipulacji znajdą sposoby na obejście zabezpieczeń. Etycznie należałoby jednak zapewnić przynajmniej procedury kontrolne: audyty niezależnych organów nadzorczych, okresowe raporty o skuteczności i przypadkach błędnej klasyfikacji, oraz możliwość interwencji człowieka w wątpliwych sytuacjach. Niepokojąca jest również perspektywa polegania przez organy ścigania wyłącznie na ocenie algorytmu – np. czy prokuratura może oskarżyć kogoś o rozpowszechnianie fałszywki wyłącznie na podstawie sygnału z automatycznego detektora? Zasada domniemania niewinności i wysoki standard dowodowy w prawie karnym nakazywałyby, by takie algorytmiczne oceny traktować jedynie pomocniczo, a nie jako rozstrzygające dowody. Etyka wymaga tu ostrożności: pomyłka algorytmu może kogoś niesłusznie napiętnować jako oszusta lub sprawcę dezinformacji, dlatego konieczna jest

[291] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, 1753-1820.

odpowiedzialność za decyzje podejmowane na podstawie AI. Odpowiedzialność ta powinna spoczywać na ludziach – operatorach systemu czy instytucji – nie zaś być rozmyta lub scedowana na „błąd komputera". Bezstronność i unikanie dyskryminacji. Podobnie jak w przypadku innych systemów AI, automatyczne narzędzia do wykrywania manipulacji mogą wykazywać stronniczość. Jeśli model był trenowany głównie na treściach w jednym języku czy z jednego kręgu kulturowego, może słabiej działać dla innych grup. Na przykład, detektor uczony na fałszywych nagraniach polityków z Europy czy USA może nie rozpoznawać subtelnych manipulacji w materiałach pochodzących z Azji czy Afryki, z uwagi na inne realia i konteksty. To rodzi ryzyko, że pewne społeczności pozostaną bardziej narażone na dezinformację niewychwyconą przez algorytm – co można uznać za nierówność informacyjną. Z drugiej strony, algorytm może też częściej jako „fałszywe" oznaczać materiały z określonych źródeł (np. niezależnych mediów) jeśli podczas uczenia przyswoił sobie jakieś nieintencjonalne uprzedzenia. Etycznie wymagane jest zatem, aby systemy te były rozwijane z dbałością o różnorodność zbiorów treningowych i testowane pod kątem biasów. Organizacje wdrażające takie narzędzia powinny prowadzić oceny skutków dla praw człowieka (ang. *Human Rights Impact Assessment*), co postulują m.in. eksperci od etyki AI[292]. Celem jest upewnienie się, że automatyczne wykrywanie manipulacji służy całemu społeczeństwu, a nie wzmacnia istniejących nierówności czy podziałów. Zaufanie społeczne i konsekwencje dla dyskursu publicznego to kolejne wyzwanie etyczne związane z automatyzacją wykrywania manipulacji medialnych. Etyczne rozważania obejmują też szerszy wpływ, jaki masowe stosowanie narzędzi wykrywających fałsz może mieć na społeczeństwo. Z jednej strony, skuteczne eliminowanie sfabrykowanych treści powinno sprzyjać ochronie prawdy w życiu publicznym – obywatele nie byliby wprowadzani w błąd przez fałszywe nagrania, a debata opierałaby się na faktach. Z drugiej strony, pojawia się ryzyko erozji zaufania: gdy ludzie uświadamiają sobie istnienie zaawansowanych technik fałszowania i ich automatycznego wykrywania, mogą zacząć kwestionować prawdziwość wszelkich materiałów, nawet autentycznych. Badania sugerują, że kontakt z *deepfake'ami* zwiększa ogólny sceptycyzm odbiorców co

[292] Komisja Europejska, A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation, 2018.

do wiarygodności mediów[293]. Równie ważne jak techniczne środki zaradcze jest zatem podnoszenie kompetencji medialnych społeczeństwa – edukacja na temat istnienia manipulacji i metod ich identyfikacji. Świadomy odbiorca, potrafiący krytycznie ocenić źródła informacji, stanowi najlepszą linię obrony przed dezinformacją. Takie podejście partycypacyjne wzmacnia odporność zbiorową na kłamstwo bez uciekania się do nadmiernej automatycznej cenzury. Działania państwa w zakresie automatycznego wykrywania manipulacji powinny podlegać publicznemu nadzorowi i być debatowane z udziałem m.in. przedstawicieli mediów, środowisk naukowych i organizacji broniących wolności w Internecie. Tylko wówczas uda się utrzymać delikatną równowagę między zapewnieniem bezpieczeństwa informacyjnego a zachowaniem otwartego, wolnego społeczeństwa. Podsumowując, automatyzacja wykrywania medialnych manipulacji to miecz obosieczny: może chronić przed realnymi szkodami wyrządzanymi przez dezinformację, ale niewłaściwie wdrażana – może sama stać się narzędziem naruszenia wartości demokratycznych. Z tego względu kluczowe jest opracowanie kodeksów etycznych i ram prawnych regulujących użycie tych technologii. Powinny one precyzować m.in. zakres dopuszczalnego monitoringu, gwarancje dla praw jednostki (prawo do odwołania, nadzór niezależny), wymogi co do bezstronności algorytmów i ochrony prywatności. Już teraz eksperci wskazują na lukę prawną w tym obszarze – brak jest wyspecjalizowanych regulacji dotyczących np. *deepfake*, wobec czego ochrony obywateli przed skutkami takich manipulacji szuka się w przepisach o ochronie wizerunku, dóbr osobistych czy danych osobowych. Konieczne jest ich uaktualnienie i uzupełnienie – tak, by prawodawstwo nadążało za rozwojem technologii[294]. Etyczne wprowadzenie systemów automatycznej detekcji wymaga zatem podejścia ostrożnościowego: krok po kroku, z pilotażami, oceną skutków i mechanizmami korekcyjnymi. Ostatecznie celem jest, aby narzędzia te działały w służbie prawdy i bezpieczeństwa, a nie kosztem praw obywatelskich. Tylko wtedy społeczeństwo obdarzy je zaufaniem i będą one mogły skutecznie spełnić swoją rolę.

[293] Zob. C. Vaccari, A. Chadwick, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News, 6(1), 1-13.

[294] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, 1753-1820.

8.3. Rola regulacji prawnych w rozwoju technologii analizy

Pomimo rozwoju zaawansowanych narzędzi technologicznych do wykrywania i weryfikacji fałszywych treści, kluczowym elementem skutecznej strategii przeciwdziałania dezinformacji pozostaje czynnik ludzki – świadomość i odporność odbiorców. Nawet najlepsze systemy analizy audiowizualnej nie spełnią swojej roli, jeśli społeczeństwo będzie podatne na manipulacje lub z góry nieufne wobec wszelkich sprostowań. Dlatego równolegle z inwestycjami w technologie detekcji państwa i organizacje międzynarodowe kładą nacisk na budowanie odporności społecznej (ang. *social resilience*) na dezinformację. Pojęcie to oznacza zdolność społeczności do przetrwania i przeciwstawienia się negatywnym skutkom rozpowszechniania fałszywych informacji – tak, by obywatele umieli rozpoznać manipulację i ograniczyć jej wpływ na swoje postawy oraz decyzje. Jak podkreśla międzynarodowa organizacja Organizacja Współpracy Gospodarczej i Rozwoju OECD[295], wzmacnianie odporności społecznej obejmuje mobilizację wszystkich sektorów społeczeństwa oraz wyposażenie jednostek w umiejętność krytycznego myślenia i rozpoznawania fałszywych treści. Chodzi zatem o uodpornienie odbiorców poprzez podnoszenie ich kompetencji informacyjnych i świadomości istnienia zagrożeń, zanim jeszcze staną się celem manipulacji. Podstawowym filarem budowania takiej odporności jest edukacja medialna i informacyjna społeczeństwa. Kompetencje medialne – obejmujące m.in. umiejętność oceny wiarygodności źródeł, rozumienia mechanizmów powstawania przekazów medialnych i świadomego korzystania z informacji – są dziś uznawane za niezbędne obywatelom w erze cyfrowej. Instytucje międzynarodowe akcentują, że edukacja w tym zakresie powinna zaczynać się już na wczesnym etapie kształcenia i trwać przez całe życie. Unia Europejska w swoich strategiach zwraca uwagę, iż kompetencje cyfrowe i medialne stanowią klucz do zwiększania odporności na manipulacje informacyjne. Oznacza to praktycznie, że programy nauczania – od szkoły podstawowej po uniwersytety trzeciego wieku – muszą uwzględniać treści związane z krytycznym odbiorem mediów, identyfikowaniem dezinformacji i świadomym

[295] OECD (ang. Organisation for Economic Co-operation and Development), Organizacja Współpracy Gospodarczej i Rozwoju), międzynarodowa organizacja założona w 1961 roku, z siedzibą w Paryżu, zrzeszająca kraje rozwinięte w celu wspierania polityk gospodarczych i społecznych.

udziałem w dyskursie publicznym. Przykładowo, Finlandia często przywoływana jest jako wzór włączania edukacji medialnej do głównego nurtu kształcenia, co postrzegane jest jako część szerszej strategii budowania odporności społeczeństwa na dezinformację. Również Organizacja Narodów Zjednoczonych ds. Edukacji, Nauki i Kultury UNESCO promuje globalnie inicjatywy z zakresu *Media and Information Literacy*[296] uznając, że umiejętność krytycznego korzystania z mediów jest warunkiem skutecznego przeciwdziałania *fakenewsom* i mowie nienawiści. We współpracy z Unią Europejską organizacja ta realizuje projekty np. *Social Media 4 Peace*[297] skoncentrowane na podnoszeniu świadomości lokalnych społeczności i szkoleniu liderów opinii w krajach narażonych na destabilizujące kampanie informacyjne. Efektem takich działań ma być społeczeństwo świadome, które potrafi samodzielnie wychwycić podejrzane sygnały w przekazach medialnych i nie ulega panice ani polaryzacji wskutek pojawienia się szokującej (fałszywej) wiadomości. Drugim istotnym elementem przeciwdziałania dezinformacji jest implementacja strategii *pre-bunking*, czyli uprzedzającego przygotowania odbiorców na potencjalne manipulacje informacyjne. Koncepcja ta, oparta na analogii do szczepienia, polega na przedstawieniu osłabionych, przykładowych fałszywych przekazów wraz z wyjaśnieniem mechanizmów ich działania, co generuje odporność na silniejsze ataki dezinformacyjne. Badania empiryczne wykazały, że takie interwencje zwiększają zdolność do krytycznego odbioru i rozpoznawania manipulacji, zmniejszając wpływ fałszywych narracji o około 20–30% w porównaniu z grupami kontrolnymi[298]. Kampanie realizowane za pomocą krótkich filmów edukacyjnych, rozpowszechnianych na platformach cyfrowych, potwierdzają efektywność *pre-bunkingowych* interwencji niezależnie od preferencji ideologicznych odbiorców. Wyniki jednoznacznie wskazują, że prewencyjna komunikacja oparta na zasadach teorii inoculacji stanowi kluczowy element budowania odporności informacyjnej

[296] https://en.unesco.org/themes/media-and-information-literacy [Dostęp 20.03.2024].

[297] Social Media 4 Peace, projekt międzynarodowy mający na celu promowanie pokoju, dialogu społecznego oraz odpowiedzialnego korzystania z mediów społecznościowych poprzez rozwój edukacji medialnej, wdrażanie strategii prewencyjnych oraz budowanie odporności społecznej na dezinformację (projekt realizowany we współpracy z partnerami międzynarodowymi, m.in. instytucjami rządowymi, organizacjami pozarządowymi i platformami cyfrowymi).

[298] Zob. S. van der Linden, A. Leiserowitz, S. Rosenthal, E. Maibach, Inoculating the public against misinformation: Evidence from two randomized controlled trials. Proceedings of the National Academy of Sciences, 114(30), 8112-8117.

społeczeństwa w dobie powszechnego rozpowszechniania dezinformacji: odbiorcy, którzy obejrzeli takie filmy, wykazywali następnie większą zdolność rozpoznania manipulacyjnych przekazów i rzadziej dawali się nabrać na dezinformację w porównaniu z grupą kontrolną. Otrzymane wyniki przedstawiona została w tabeli 7.

Tabela 7. Wyniki eksperymentalne – efekty prewencyjnej komunikacji opartej na teorii inoculacji

Warunki eksperymentalne	Średnia (przed) [%]	Średnia (po) [%]	Różnica (po – przed) [%] (SE)	Cohen's D (w porównaniu z grupą kontrolną)
Grupa kontrolna (n = 360)	72,18	72,53	0,35 (0,36)	–
Konsensus (CT) (n = 338)	70,58	90,30	19,72 (1,17)	1,23
Countermessage (CM) (n = 392)	72,04	63,05	–8,99 (1.31)	0,48
CT	CM (n = 352)	73,48	72,99	–0,51 (1,39)
CT + ogólna inoculacja	CM (n = 363)	73,29	79,76	6,47 (1,32)
CT + szczegółowa inoculacja	CM (n = 362)	71,23	83,94	12,71 (1,17)

Źródło: S. van der Linden, A. Leiserowitz, S. Rosenthal, E. Maibach, Inoculating the public against misinformation: Evidence from two randomized controlled trials. Proceedings of the National Academy of Sciences, 114(30), 2017, s. 8112-8117.

Co istotne, efekt ten obserwowano niezależnie od preferencji politycznych widzów czy ich poglądów, co sugeruje, że *pre-bunking* może działać na szeroką skalę w zróżnicowanych populacjach. Idąc za tymi badaniami, niektóre rządy i organizacje zaczęły eksperymentować z wdrażaniem strategii *pre-bunking* w praktyce. Departament Stanu USA w opracowanej analizie[299] wskazał, że kampanie uprzedzające (*prebunking*) powinny stanowić istotny element walki z dezinformacją, obok promowania pozytywnych, opartych na faktach narracji. Z kolei Agencja Stanów Zjednoczonych ds. Rozwoju Międzynarodowego USAID[300] we współpracy z partnerami uruchomiła

[299] Departament Stanu USA, Strategic Communication Report: Prebunking as a Tool to Counter Misinformation, 2022.

[300] USAID (ang. United States Agency for International Development), Agencja Stanów Zjednoczonych ds. Rozwoju Międzynarodowego, założona w 1961 roku, jest niezależną agencją rządową odpowiedzialną za udzielanie pomocy rozwojowej i humanitarnej na całym świecie, wspierając programy mające na celu redukcję ubóstwa, poprawę zdrowia i edukacji oraz rozwój gospodarczy i społeczny.

w 2023 roku serię filmów prewencyjnych w Europie Środkowo-Wschodniej[301], wymierzonych w narracje antyuchodźcze. Tego rodzaju inicjatywy wskazują na rodzący się konsensus, że lepiej zapobiegać, niż korygować – uprzednio uświadomiony odbiorca jest mniej podatny na fałsz, dzięki czemu szkodliwe treści nie zdobywają tak dużego zasięgu jak w społeczeństwie nieprzygotowanym. Oczywiście, psychologiczne uwarunkowania odbioru informacji sprawiają, że nie istnieje proste rozwiązanie gwarantujące całkowitą odporność społeczności na kłamstwo. Ludzkie mechanizmy poznawcze cechują się skłonnością do upraszczania i filtrowania informacji – często wierzymy w te przekazy, które odpowiadają wcześniejszym przekonaniom lub emocjonalnym potrzebom. Zjawisko to znane jako stronniczość potwierdzania powoduje, że nawet w obliczu obiektywnych dowodów obalających jakąś fałszywą historię, część odbiorców może trwać przy swoim przekonaniu, jeśli kłamstwo było zgodne z ich światopoglądem. Ponadto ładunek emocjonalny dezinformacji – sensacyjne, szokujące treści – sprawia, że rozprzestrzenia się ona szybciej niż żmudne sprostowania. Walka z tym wymaga nie tylko edukowania o faktach, ale także uświadamiania społeczeństwu istnienia mechanizmów psychologicznych wykorzystywanych przez manipulatorów. Dlatego programy zwiększania odporności coraz częściej łączą elementy czysto informacyjne (np. kampanie *fact-checkingowe*) z komponentami uczącymi rozpoznawać techniki perswazji, odwoływania się do uprzedzeń czy wywoływania strachu. Istotne jest również budowanie zaufania społeczeństwa do zweryfikowanych źródeł informacji – obywatele, którzy ufają instytucjom publicznym i profesjonalnym mediom, rzadziej dają wiarę niesprawdzonym rewelacjom z przypadkowych źródeł. W tym kontekście rządy i organizacje pozarządowe prowadzą działania na rzecz przejrzystości komunikacji i szybkiego reagowania na pojawiające się fałsze, tak aby nie pozostawiać próżni informacyjnej w której mogłaby rozwinąć się dezinformacja. NATO w swojej strategii podkreśla znaczenie proaktywnej komunikacji ze strony instytucji – polegającej na szybkim dostarczaniu społeczeństwu rzetelnych informacji, co utrudnia rozsiewanie plotek i manipulacji. Takie wyprzedzające informowanie nie tylko niweluje wpływ konkretnych nieprawdziwych przekazów, ale z czasem wzmacnia też odporność poznawczą obywateli: widząc, że pewne narracje są systematycznie dementowane

[301] USAID, Prebuwomenking for Resilience: A Campaign Against Migrant-Targeted Disinformation in Central Eastern Europe, 2023.

przez wiarygodne źródła, uczą się oni podchodzić z rezerwą do sensacyjnych doniesień niepopartych oficjalnym potwierdzeniem. Podsumowując, odporność społeczna stanowi nieodzowny komponent systemu bezpieczeństwa informacyjnego państwa w dobie audiowizualnej dezinformacji. Osiągnięcie jej wymaga długofalowych inwestycji w edukację i świadomość medialną, a także wdrażania innowacyjnych podejść prewencyjnych, takich jak *pre-bunking* oparty na teorii szczepień informacyjnych. Rządy, organizacje międzynarodowe i społeczeństwo obywatelskie muszą współdziałać w tych wysiłkach – żadne pojedyncze narzędzie (technologiczne czy edukacyjne) nie zapewni pełnej ochrony. Dopiero kombinacja skutecznych technik detekcji z wysoką świadomością społeczną i zaufaniem do prawdziwych informacji pozwala ograniczyć wpływ wrogich operacji informacyjnych. W literaturze przedmiotu coraz częściej podkreśla się, że sukces w walce z dezinformacją nie będzie mierzony wyłącznie liczbą wykrytych *deepfake'ów*, ale przede wszystkim odpornością społeczeństwa na ich oddziaływanie – zdolnością do zachowania spokoju, krytycznego myślenia i jedności wobec prób manipulacji narracją publiczną. W ten sposób technologia i czynnik ludzki łączą się w komplementarną strategię ochrony przestrzeni informacyjnej demokratycznych państw.

8.4. Rozwój AI w analizie treści audiowizualnych i jego konsekwencji

Intensywny rozwój sztucznej inteligencji w ostatnich dwóch dekadach radykalnie zmienił możliwości analizy treści audiowizualnych. Tradycyjne metody, oparte niegdyś na prostych algorytmach detekcji ruchu czy statycznych regułach (np. wykrywających przekroczenie wyznaczonej linii w obrazie z kamery), ustępują miejsca systemom uczącym się na przykładach. Uczenie głębokie (ang. *deep learning*) sprawiło, że komputery zaczęły rozpoznawać obrazy i dźwięki z precyzją dotąd nieosiągalną. Kamieniem milowym była tu rewolucja zapoczątkowana około 2012 roku – zastosowanie głębokich sieci neuronowych do klasyfikacji obrazów (*AlexNet)* pokazało skokowy wzrost skuteczności rozpoznawania obiektów na zdjęciach. W ślad za tym przyszły rozwiązania dedykowane wideo: modele potrafiące śledzić wiele obiektów naraz, rozpoznawać twarze w tłumie czy

identyfikować złożone aktywności na nagraniach. Skok wydajnościowy szedł w parze z gwałtownym obniżeniem kosztów obliczeń – potężne GPU i usługi chmurowe umożliwiły trenowanie modeli na olbrzymich zbiorach nagrań w rozsądnym czasie.

W efekcie w połowie lat 2010. nastąpiła demokratyzacja technologii analizy audiowizualnej: zaczęły z niej korzystać nie tylko wyspecjalizowane agencje rządowe, ale też organy ścigania na poziomie lokalnym, a nawet prywatne firmy. Pozytywne konsekwencje. Rozwój AI wyposażył służby bezpieczeństwa w zupełnie nowe narzędzia. Przede wszystkim poprawiła się skuteczność i szybkość przetwarzania informacji. Zadania, które dawniej wymagały wielu godzin pracy analityków (np. ręcznego przeglądania zapisu z kamer monitoringu w poszukiwaniu określonej osoby), dziś mogą być w dużej mierze zautomatyzowane. Systemy rozpoznawania twarzy potrafią w ciągu sekund przeszukać bazy danych i zidentyfikować podejrzanego uchwyconego przez kamerę uliczną. Głośnym przykładem jest zatrzymanie poszukiwanego przestępcy na koncercie w Chinach – wśród dziesiątek tysięcy obecnych, dzięki kamerom i AI, policja wychwyciła jedną twarz zgodną z bazą poszukiwanych[302]. Choć taki sukces budzi również pytania o nadzór, pokazuje niewątpliwą skuteczność nowej technologii. W miastach zachodnich również testuje się systemy automatycznego wykrywania zagrożeń – np. bójek, ataków czy wypadków – na żywo w strumieniu wideo, by natychmiast alarmować odpowiednie służby. AI jest też stosowana do analizy nagrań z dronów patrolujących granice czy tereny przygraniczne, co pozwala szybciej wykrywać próby przemytu lub nielegalnego przekroczenia granicy. Istotnym polem zastosowań stała się identyfikacja i śledzenie osób podejrzanych o terroryzm lub poważne przestępstwa. Dzięki połączeniu analizy obrazu z bazami danych (listy poszukiwanych, rejestry migrantów itp.), służby mogą efektywniej monitorować przemieszczanie się podejrzanych. W skali państwa oznacza to potencjalne zwiększenie bezpieczeństwa – łatwiej ująć zbiega czy zapobiec aktowi przemocy. Innym przykładem pozytywnej konsekwencji rozwoju AI jest usprawnienie analizy materiałów dowodowych.

Algorytmy potrafią automatycznie wychwycić moment strzału na nagraniu audio-wideo, zliczać strzały, określać kaliber broni na podstawie dźwięku – to cenne informacje dla dochodzeń. W dziedzinie pornografii

[302] ACLU, Williams v. City of Detroit (skarga dotycząca niesłusznego aresztowania w oparciu o algorytm rozpoznawania twarzy), American Civil Liberties Union, 2020.

dziecięcej wyspecjalizowane systemy klasyfikują filmy i zdjęcia, wspierając śledczych w odsiewaniu ogromnych ilości danych i kierując ich uwagę na najbardziej podejrzane treści (co przyspiesza identyfikację ofiar i sprawców). Również w wywiadzie i kontrwywiadzie postępy AI przynoszą korzyści: techniki rozpoznawania mowy i transkrypcji automatycznej umożliwiają szybkie przeanalizowanie przechwyconych rozmów czy nagrań, np. wykrycie słów kluczowych wskazujących na planowany zamach. Ważną korzyścią jest pojawienie się narzędzi do walki z dezinformacją. Ten sam rozwój, który umożliwił tworzenie *deepfake'ów*, przyniósł też metody ich detekcji. Zaawansowane algorytmy są w stanie analizować materiały medialne pod kątem spójności obrazu i dźwięku, wykrywać ślady modyfikacji lub nienaturalne artefakty. Dzięki temu instytucje państwowe mogą szybciej dementować fałszywe informacje. Przykładowo, materiały propagandowe organizacji terrorystycznych mogą być automatycznie wykrywane i usuwane z platform internetowych – opracowane we współpracy rządowej systemy potrafią rozpoznawać ponad 90% treści ISIS w sieci, zanim ktokolwiek je zgłosi. To realnie utrudnia terrorystom szerzenie przekazu i utrzymywanie kanałów rekrutacji. AI pomaga również w analizie jawnych przekazów medialnych pod kątem wrogiej propagandy – w wyszukiwaniu powtarzających się narracji dezinformacyjnych, identyfikacji zautomatyzowanych kont (*botów*) rozpowszechniających fałsz. W ten sposób służby mogą wcześniej wykryć zorganizowane kampanie informacyjne wymierzone w państwo i podjąć kontrdziałania. Negatywne konsekwencje i zagrożenia to kolejne wyzwanie etyczne.

Dynamiczny rozwój AI, obok oczywistych korzyści, pociąga za sobą także szereg negatywnych konsekwencji – zarówno nowych zagrożeń, jak i niezamierzonych skutków ubocznych. Paradoksalnie, te same algorytmy, które wzmacniają ochronę, mogą zostać wykorzystane przez przeciwników do ataku. *Deepfake* i inne syntetyczne media stały się dostępne nie tylko rządom, ale również grupom przestępczym czy nawet pojedynczym hakerom. To otwiera zupełnie nowy front wojny informacyjnej. W sferze międzynarodowej istnieje obawa, że państwa wrogie lub organizacje terrorystyczne będą używać fabrykowanych nagrań do siania chaosu politycznego. Już w 2019 r. odnotowano przypadki użycia *deepfake* do celów politycznych – w Gabonie pojawiło się podejrzenie, że przemówienie prezydenta zostało zmanipulowane komputerowo, co stało się jedną z przesłanek zamachu stanu. W Malezji

w tym samym roku krążyło rzekome nagranie kompromitujące jednego z polityków, które również uznano za potencjalnie wygenerowane cyfrowo. Choć te sytuacje pozostają kontrowersyjne, wskazują jasno, iż *deepfake* może być narzędziem rozgrywek politycznych na najwyższym szczeblu. W 2022 roku miało miejsce prawdopodobnie pierwsze w historii wykorzystanie *deepfake* w trakcie pełnoskalowego konfliktu zbrojnego – w Internecie pojawiło się fałszywe wideo przedstawiające prezydenta Ukrainy Wołodymyra Zełenskiego wzywającego swoje wojska do złożenia broni. Nagranie to zostało szybko zdemaskowane i usunięte z głównych platform, jednak przez krótki czas zdołało zaistnieć, m.in. wyświetlono je na zhakowanej stronie telewizji Ukraine 24. Incydent ten ukazał bardzo realne zagrożenie: przeciwnik może próbować użyć syntetycznych mediów, by wywołać panikę, dezorientację lub podkopać morale społeczeństwa i armii. Analitycy określili to zdarzenie jako „wierzchołek góry lodowej" nowych form wojny informacyjnej, sugerując, że podobnych prób będzie więcej. Na poziomie wewnętrznym państw rozwój AI również niesie skutki ambiwalentne. Przestępcy i oszuści zyskują potężne możliwości. Pojawiły się już przypadki wykorzystania syntezatora mowy do podszycia się pod czyjś głos i wyłudzenia dużych sum pieniędzy – w 2019 r. zanotowano oszustwo, w którym cyberprzestępcy użyli *deepfake* audio imitującego głos prezesa firmy, by nakłonić pracownika do wykonania nieautoryzowanego przelewu. Tego rodzaju incydenty pokazują, że AI staje się bronią w rękach grup przestępczych, umożliwiając tzw. socjotechniczne ataki na niespotykaną dotąd skalę.

Rozwój technik generatywnych może także ułatwiać działania w cyberprzestępczości – np. automatyczne tworzenie przekonujących *fake newsów*, masek biometrycznych do oszukiwania systemów uwierzytelniania w tym bankowych rozpoznań głosu czy systemów kontroli granicznej lub generowanie materiałów szantażujących. Erozja zaufania społecznego to kolejna konsekwencja rozwoju AI w analizie mediów. Jak zauważono wcześniej, im bardziej przekonujące fałszywe treści potrafi stworzyć technologia, tym bardziej maleje pewność co do autentyczności przekazów. Społeczeństwo, bombardowane informacjami o *deepfake'ach,* może popaść w cynizm informacyjny – przekonanie, że „nic nie jest prawdziwe". To zjawisko zostało określone mianem efektu „epistemicznego chaosu" lub właśnie dywidendy kłamcy, gdzie już samo istnienie zaawansowanych fałszerstw

podważa wiarygodność prawdziwych nagrań[303]. Konsekwencje dla państwa są poważne: spada zaufanie do mediów, do komunikatów władz, a nawet do dowodów przedstawianych w procesach sądowych. Rozwój AI wymusza zatem dostosowanie procedur i standardów – np. sądy muszą wprowadzać dodatkowe ekspertyzy weryfikujące autentyczność materiału wideo przed uznaniem go za dowód, co wydłuża i komplikuje postępowania. Służby prasowe instytucji publicznych zmuszone są do szybkiego reagowania na potencjalne *fake newsy*, a nawet prewencyjnego „przygotowywania" odbiorców na to, że mogą pojawić się takie fałszywki, co obserwowano np. w Ukrainie, gdzie już przed incydentem z *deepfake* Zelenskiego ostrzegano społeczeństwo przed taką ewentualnością.

Jeszcze innym efektem ubocznym upowszechnienia systemów AI jest zmiana charakteru działalności służb i potencjalne poleganie na algorytmach w stopniu, który może być niebezpieczny. Istnieje ryzyko, że zbyt duże zaufanie do automatycznej analizy doprowadzi do zaniedbania tradycyjnych metod i ludzkiej czujności. Przykładowo, jeżeli system oceni, że dany materiał jest autentyczny, a w rzeczywistości okaże się on sprytną fałszywką niewykrytą przez algorytm, ludzie mogą zostać wprowadzeni w błąd z powodu nadmiernej wiary w technologię. Podobnie, systemy predykcyjnej analizy zachowań (ang. *predictive policing*) analizują ogrom danych, sugerując priorytety działań policji – co może prowadzić do nadużyć lub błędów, jeśli algorytm się myli lub jest stronniczy. Już odnotowano przypadki błędnej identyfikacji przez system rozpoznawania twarzy, skutkujące niesłusznym aresztowaniem niewinnej osoby, co wywołało publiczną krytykę i pokazało, że nadmierna wiara w nieomylność AI może naruszyć prawa obywateli. Konsekwencją rozwoju AI jest więc konieczność utrzymania ludzkiego czynnika w pętli decyzyjnej oraz ciągłego weryfikowania poprawności i sprawiedliwości algorytmów[304]. Wymaga to dodatkowych szkoleń personelu i powołania wyspecjalizowanych komórek nadzorczych oceniających działanie tych systemów w praktyce. Rozwój AI w analizie treści audiowizualnych przypomina zatem wyścig zbrojeń – każda innowacja

[303] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, 1753-1820.

[304] Komisja Europejska, A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation, 2018.

przynosi zarówno nowe możliwości, jak i reakcje drugiej strony, starającej się ją obejść lub wykorzystać w niecny sposób.

Państwo musi mierzyć się z rosnącą ilością danych do analizy oraz coraz bardziej wyrafinowanymi metodami ich fałszowania. W odpowiedzi pojawiają się inicjatywy mające przywrócić pewność co do źródeł informacji, jak np. oznaczanie oryginalnych nagrań cyfrowymi znacznikami pochodzenia (projekty typu *Content Authenticity Initiative*) czy rozwój standardów weryfikacji multimediów na poziomie całego przemysłu. To jednak dopiero wschodzące technologie i standardy, których upowszechnienie może potrwać. Tymczasem organy państwa już teraz muszą przystosować swoje procedury do nowej rzeczywistości – zwiększać nakłady na własne systemy AI, aby nie pozostać w tyle za adwersarzami, a zarazem tworzyć bezpieczniki prawne i organizacyjne, by nadmiernie nie zaufać maszynom tam, gdzie stawką są ludzkie losy. Wszystko to stanowi ogromne wyzwanie strategiczne. Konsekwencje rozwoju AI są bowiem wielowymiarowe: to nie tylko usprawnienie działań operacyjnych, ale i konieczność transformacji całego ekosystemu bezpieczeństwa – od szkoleń kadr, przez zmiany legislacyjne, po budowę międzynarodowej współpracy w zakresie standardów i wymiany informacji o zagrożeniach. Te właśnie aspekty – adaptacji systemów bezpieczeństwa państwa do nowych realiów – zostaną omówione w kolejnym podrozdziale, zawierającym rekomendacje dotyczące polityki w tym obszarze.

8.5. Interdyscyplinarność w rozwoju narzędzi do wykrywania dezinformacji

Rozpoznawanie dezinformacji, zwłaszcza w formie treści audiowizualnych, stanowi jedno z największych wyzwań w obszarze bezpieczeństwa informacyjnego XXI wieku. Wraz z gwałtownym rozwojem technologii generatywnych, takich jak deepfake czy syntetyczny dźwięk, potrzeba tworzenia systemów zdolnych do wykrywania fałszywych treści staje się coraz bardziej paląca. Jednak, jak pokazuje doświadczenie ostatnich lat, skuteczne narzędzia wykrywające manipulacje nie mogą być wyłącznie efektem pracy informatyków czy inżynierów. Wymagają one głębokiego zakorzenienia w wiedzy interdyscyplinarnej, łączącej elementy nauk ścisłych, społecznych i humanistycznych.

Technologie analityczne oparte na sztucznej inteligencji (AI) i uczeniu maszynowym (ML) są niezbędne do przetwarzania ogromnych ilości danych[305], identyfikowania powtarzalnych wzorców manipulacji oraz weryfikacji autentyczności materiałów audiowizualnych. Jednak skuteczność tych narzędzi w dużej mierze zależy od sposobu, w jaki są projektowane, trenowane i testowane. Do właściwego „nauczenia" systemu konieczne jest dostarczenie mu nie tylko danych, ale także zrozumienia, czym w istocie jest manipulacja i jak rozpoznaje ją człowiek. W tym miejscu kluczową rolę odgrywa psychologia poznawcza, która bada mechanizmy percepcji i przetwarzania informacji wizualnych, a także sposoby, w jakie ludzie reagują na bodźce emocjonalne zawarte w obrazie i dźwięku.

Nie mniej istotna jest rola lingwistyki i semiotyki – dziedzin, które pozwalają analizować strukturę komunikatu, jego warstwę językową, symbolikę oraz kontekst kulturowy. Treści dezinformacyjne często posługują się językiem nacechowanym emocjonalnie, powtarzającymi się frazami retorycznymi, a także strukturami gramatycznymi i stylistycznymi charakterystycznymi dla określonych środowisk ideologicznych. Tylko dzięki integracji wiedzy lingwistycznej możliwe jest stworzenie algorytmów wykrywających subtelne różnice między neutralną a zmanipulowaną treścią – zwłaszcza w przypadku analiz wielojęzycznych i transkulturowych.

Kolejnym istotnym komponentem interdyscyplinarności jest medioznawstwo, które wnosi do procesu analitycznego wiedzę na temat sposobów funkcjonowania mediów, ich wpływu na opinię publiczną, charakteru przekazów audiowizualnych oraz dynamiki mediów społecznościowych. Znajomość logiki algorytmicznej platform cyfrowych, strategii narracyjnych oraz historii wojen informacyjnych pozwala nie tylko rozpoznawać dezinformację, ale również przewidywać, w jakich warunkach i z jakim celem będzie się ona pojawiać. Medioznawcy i socjologowie informacji są również w stanie wskazać typowe wzorce zachowań odbiorców w reakcji na określone formy przekazu wizualnego – co stanowi cenny materiał przy kalibracji systemów predykcyjnych.

Nie można również pominąć wkładu nauk o bezpieczeństwie oraz studiów strategicznych, które dostarczają ramy analityczne do rozumienia dezinformacji jako elementu działań hybrydowych, operacji wpływu czy

[305] J. Nowicka, Y. Pauliuchuk, Z. Ciekanowski, B. Falda, K. Sikora, The Use of Data Analytics in Human Resource Management, European Research Studies Journal, XXVII(2), 205.

zagrożeń asymetrycznych. Zrozumienie mechanizmów prowadzenia kampanii dezinformacyjnych przez podmioty państwowe lub organizacje niepaństwowe pozwala projektować narzędzia wykrywające nie tylko fałszywe informacje, ale także całe struktury narracyjne, które mogą destabilizować system polityczny, osłabiać morale społeczne czy ingerować w procesy demokratyczne. Ekspertyza w dziedzinie bezpieczeństwa pozwala też ocenić znaczenie danej treści nie tylko z punktu widzenia jej autentyczności, ale także potencjalnego wpływu społeczno-politycznego.

Interdyscyplinarne podejście obejmuje również aspekt etyczny i prawny. Specjaliści z zakresu prawa mediów, prawa do prywatności oraz ochrony danych osobowych odgrywają ważną rolę w ocenie, jakie techniki analizy treści są dopuszczalne, a które mogą naruszać prawa jednostki. Współpraca informatyków z prawnikami, etykami technologii i socjologami cyfrowymi pozwala tworzyć narzędzia zgodne z zasadami odpowiedzialności społecznej, przejrzystości i sprawiedliwości informacyjnej. To szczególnie istotne w kontekście ryzyka nadinterpretacji danych, błędów algorytmicznych czy nieuprawnionej automatyzacji decyzji o charakterze społeczno-politycznym.

Wreszcie, interdyscyplinarność jest niezbędna także w zakresie wdrażania i testowania nowych narzędzi w realnym środowisku społecznym[306]. Ostateczna skuteczność systemów wykrywających dezinformację nie zależy bowiem wyłącznie od ich precyzji technicznej, ale także od stopnia, w jakim zostaną one zaakceptowane przez instytucje publiczne, sektor prywatny oraz społeczeństwo. Tylko rozwiązania opracowane we współpracy z użytkownikami końcowymi – dziennikarzami, urzędnikami, nauczycielami, odbiorcami informacji – mają szansę rzeczywiście wspierać bezpieczeństwo informacyjne, a nie stwarzać nowe ryzyka związane z automatyzacją i brakiem przejrzystości.

Podsumowując, skuteczna walka z dezinformacją w dobie obrazu nie może opierać się wyłącznie na rozwoju technologii detekcji. Wymaga ona ścisłej współpracy specjalistów z różnych dziedzin, którzy razem projektują systemy zdolne nie tylko do rozpoznawania fałszu, ale także rozumienia jego kontekstu, intencji i wpływu społecznego. Tylko takie podejście zapewni powstanie narzędzi nie tylko inteligentnych, ale również odpowiedzialnych, humanistycznych i skutecznych.

[306] G. Mazurek, Sztuczna inteligencja, prawo i etyka, Krytyka Prawa, 1, 2023, s. 8-9.

IX

Wnioski, rekomendacje i kierunki dalszych działań i badań

9.1. Rekomendacje dla polityki bezpieczeństwa państwa

Analiza przedstawionych wyżej wyzwań technologicznych, aspektów etycznych oraz konsekwencji rozwoju AI pozwala sformułować szereg rekomendacji dla polityki bezpieczeństwa państwa. Mają one na celu zrównoważenie korzyści płynących z nowych narzędzi analizy treści audiowizualnych z koniecznością ochrony praw obywateli i zapewnienia odporności państwa na potencjalne zagrożenia. Rekomendacje te można ująć w kilku obszarach: regulacji prawnych, standardów etycznych, rozwoju kompetencji i infrastruktury technicznej, współpracy międzynarodowej oraz edukacji społecznej.

1. Dostosowanie i uzupełnienie ram prawnych. Państwo powinno pilnie zaktualizować swoje ustawodawstwo, tak by odpowiadało na zjawisko głębokich fałszerstw i automatycznej analizy treści. Konieczne jest zdefiniowanie w przepisach pojęć takich jak fałszerstwo audiowizualne (np. *deepfake*) oraz wprowadzenie jasnych sankcji za ich złośliwe wykorzystywanie. Obecnie ściganie sprawców bazuje na artykułach o zniesławieniu, oszustwie czy naruszeniu wizerunku, co bywa niewystarczające wobec nowych form ataków informacyjnych. Prawo powinno przewidywać kary za tworzenie i dystrybucję nieuprawnionych *deepfake'ów* wyrządzających szkodę jednostkom oraz za użycie ich w celach destabilizacji społecznej. Należy rozważyć wprowadzenie szczególnych typów przestępstw, np. fałszerstwa medialnego w kontekście bezpieczeństwa narodowego, z surowszymi sankcjami, gdy sprawcą jest podmiot działający na rzecz obcego wywiadu lub organizacji

terrorystycznej. Ponadto, ustawodawca powinien nałożyć na platformy internetowe obowiązki w zakresie reagowania na oczywiste fałszywe treści zagrażające bezpieczeństwu. Może to obejmować wymóg niezwłocznego informowania odpowiednich służb o pojawieniu się *deepfake* wymierzonego w infrastrukturę krytyczną lub porządek publiczny oraz zapewnienie dostępu do danych pozwalających namierzyć twórcę takiej treści zgodnie z zasadami ochrony danych. Wreszcie, prawo powinno ułatwiać ofiarom manipulacji medialnych dochodzenie swoich praw – np. przez wprowadzenie szybkiej ścieżki nakazowej do usunięcia *deepfake* naruszającego dobra osobiste i zabezpieczenie materiału dowodowego, zanim rozpowszechni się on *viralowo*. Równolegle, istotne jest dostosowanie przepisów dotyczących dowodów cyfrowych i procedur sądowych. Organy ścigania i wymiar sprawiedliwości muszą dysponować narzędziami prawnymi do weryfikacji autentyczności nagrań – np. możliwość zarządzenia specjalistycznej ekspertyzy technicznej w toku postępowania, z odwróceniem ciężaru dowodu w sytuacji uzasadnionego podejrzenia manipulacji[307]. W kontekście prawa międzynarodowego Polska i inne państwa powinny inicjować dialog nad uregulowaniem użycia zaawansowanych fałszerstw w konfliktach i operacjach wpływu. Warto rozważyć, czy wykorzystanie *deepfake* w celach destabilizacji innego kraju nie powinno być traktowane jako naruszenie prawa międzynarodowego. Choć trudno wypracować globalny konsensus, już sama dyskusja w ONZ czy NATO może zwiększyć świadomość i prowadzić do stworzenia *soft law* – niewiążących wytycznych – w tym zakresie.

2. Standardy etyczne i nadzór nad stosowaniem AI. Wdrażając systemy automatycznej analizy treści, państwo powinno kierować się przejrzystym kodeksem etycznym, zgodnym z zasadami państwa prawa. Niezbędne jest ustanowienie mechanizmów kontroli i równowagi (ang. *checks and balances*) nad użyciem takich narzędzi przez służby. Rekomenduje się utworzenie niezależnego organu lub funkcji, przykładem może być Inspektor Algorytmiczny przy Rzeczniku Praw Obywatelskich lub Najwyższej Izbie Kontroli, który monitorowałby wykorzystanie AI w obszarze bezpieczeństwa i regularnie raportował o potencjalnych nadużyciach, błędach czy *biasach*[308]. Organy

[307] Zob. R. Chesney, D. Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, 1753-1820.

[308] Komisja Europejska, A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation, 2018.

korzystające z analizy audiowizualnej opartej na SI powinny obowiązkowo prowadzić oceny skutków dla prywatności i praw człowieka przed wdrożeniem danego systemu. Takie podejście pozwoli zidentyfikować zawczasu ryzyka między innymi dyskryminację grup w wyniku działania algorytmu i wprowadzić środki zaradcze. Konieczne jest również sformułowanie jasnych reguł operacyjnych. Przykładem może być fakt, iż algorytmicznie wygenerowane oceny i alerty nie mogą stanowić wyłącznej podstawy decyzji o zatrzymaniu obywatela czy wszczęciu postępowania – zawsze musi nastąpić niezależna weryfikacja przez człowieka (zasada *human-in-the--loop*). Należy wprowadzić protokoły dokumentowania procesu analizy – tak, aby w razie sporu można było zbadać, jak doszło do danej rekomendacji AI. Choć pełna wyjaśnialność skomplikowanych sieci neuronowych bywa technicznie trudna, minimum powinno obejmować zachowanie logów działania systemu, wersji modelu, metryk jego skuteczności i znanych ograniczeń. Wszelkie aktualizacje algorytmów w krytycznych zastosowaniach muszą być notyfikowane odpowiednim organom nadzorczym. Można tu wzorować się na mechanizmach certyfikacji oprogramowania medycznego czy systemów automatyzujących ruch lotniczy – AI wpływająca na bezpieczeństwo publiczne powinna przechodzić podobne procedury akredytacji pod kątem zgodności z normami.

3. Rozwój zdolności technicznych i współpraca z sektorem prywatnym. Państwo powinno inwestować w rozwój własnych kompetencji i infrastruktury w zakresie AI, tak aby nie być całkowicie zależnym od zewnętrznych dostawców technologii[309]. Rekomenduje się utworzenie specjalistycznych jednostek badawczo-operacyjnych jak laboratorium analizy mediów cyfrowych przy ABW lub policji, wyposażonych w najnowsze narzędzia do wykrywania manipulacji i prowadzenia analizy forensycznej nagrań. Jednostki te mogłyby korzystać z wiedzy akademickiej – warto zacieśnić współpracę pomiędzy służbami a ośrodkami naukowymi specjalizującymi się w przetwarzaniu obrazu i dźwięku. Programy grantowe finansowane ze środków publicznych powinny wspierać badania nad skuteczniejszym wykrywaniem *deepfake*, metodami uwierzytelniania materiałów znakowaniem treści w momencie rejestracji oraz nad zabezpieczeniami przed atakami na systemy AI odpornością na przykłady *adversarialne*. Współpraca

[309] Zob. R. Chesney, D. Citron*, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, California Law Review, 107, 1753-1820.

z sektorem prywatnym jest równie istotna. Większość platform, gdzie rozpowszechniane są treści (Facebook, Twitter, YouTube), jest w rękach firm technologicznych – państwo powinno więc aktywnie z nimi współdziałać. Rekomenduje się wypracowanie protokołów szybkiej wymiany informacji. W sytuacji, gdy służby odkryją kampanię dezinformacyjną z użyciem *deepfake*, platformy powinny być w stanie błyskawicznie otrzymać stosowny raport i podjąć działania takie jak oznaczenie treści, ograniczenie zasięgu, czy ostrzeżenie użytkowników. W tym celu można ustanowić stałe centrum reagowania kryzysowego z udziałem przedstawicieli państwa i dużych platform – analogicznie do istniejących zespołów ds. cyberbezpieczeństwa. Ponadto, państwo może zachęcać sektor prywatny do dobrowolnego przyjmowania zobowiązań w ramach kodeksów postępowania. Przykładem jest unijny Kodeks postępowania w zakresie dezinformacji z 2018 r., w którym firmy zadeklarowały m.in. zwalczanie fałszywych kont i transparentność reklam politycznych (European Commission, 2018). W najnowszej aktualizacji tego kodeksu (2022) uwzględniono też problem deepfake i zapowiedziano oznaczanie treści syntetycznych. Rząd powinien monitorować realizację tych zobowiązań i w razie potrzeby dążyć do ich usankcjonowania w prawie (np. poprzez krajowe regulacje uzupełniające unijne rozporządzenia DSA i AI Act). Co więcej, zaleca się promowanie standardów technicznych zapewniających integralność mediów. Państwo może wspierać wdrażanie inicjatyw takich jak Content Authenticity Initiative – np. poprzez stosowanie podpisów cyfrowych i metadanych potwierdzających autentyczność materiałów wytwarzanych przez instytucje publiczne. Gdy obywatele będą mogli łatwo zweryfikować, że komunikat rzeczywiście pochodzi od urzędu (a nie jest fałszerstwem), zmniejszy się podatność na dezinformację. W tym duchu w projekcie unijnego Aktu o SI przewidziano obowiązek oznaczania generowanych sztucznie obrazów lub nagrań[310]. Polityka państwa powinna zmierzać do jak najszerszego upowszechnienia takich dobrych praktyk również poza sektor publiczny – poprzez zachęty dla mediów i twórców internetowych do dobrowolnego dołączania informacji o pochodzeniu i edycjach publikowanych treści.

4. Współpraca międzynarodowa i wymiana informacji. Zagrożenia związane z manipulacjami audiowizualnymi mają charakter transgraniczny

[310] European Commission, Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), European Commission, 2021.

– filmy i nagrania rozprzestrzeniają się globalnie, a ataki informacyjne często są inspirowane lub finansowane spoza kraju ofiary. Współczesna skala i transgraniczny charakter zagrożeń informacyjnych sprawiają, że żadne państwo nie jest w stanie samodzielnie zapewnić pełnej ochrony swojej przestrzeni informacyjnej. W związku z tym wskazane jest intensyfikowanie współpracy międzynarodowej – zarówno w ramach sojuszy polityczno-obronnych, takich jak NATO czy Unia Europejska, jak i na płaszczyźnie dwustronnej z kluczowymi partnerami strategicznymi. W ramach NATO warto rozważyć utworzenie centrum doskonalenia (ang. *Centre of Excellence*) poświęconego zagrożeniom informacyjnym i *deepfake* – podobnie jak istnieją centra NATO ds. cyberobrony czy komunikacji strategicznej. Takie centrum mogłoby koordynować badania nad nowymi metodami detekcji oraz prowadzić ćwiczenia symulacyjne między państwami, przygotowujące na scenariusz masowego ataku z użyciem *fake news* i *deepfake*. Na poziomie Unii Europejskiej należy dążyć do harmonizacji podejścia – wspólne standardy prawne powinny być uzupełniane wspólnymi mechanizmami operacyjnymi. Dobrym krokiem było utworzenie unijnego Systemu Wczesnego Ostrzegania przed dezinformacją, który pozwala krajom członkowskim dzielić się informacjami o kampaniach dezinformacyjnych prowadzonych przez podmioty zagraniczne[311]. Należy rozszerzyć jego działanie o komponent związany z wykrywaniem zaawansowanych fałszerstw – stworzyć paneuropejski rejestr zidentyfikowanych *deepfake'ów* zaangażowanych w szkodliwe działania, dostępny dla służb wszystkich państw UE. Ponadto, Polska powinna aktywnie uczestniczyć w inicjatywach międzynarodowych wypracowujących normy i dobre praktyki w dziedzinie AI. UNESCO w 2021 r. przyjęło Rekomendację w sprawie etyki AI, w której podkreślono znaczenie globalnej współpracy i ochrony praw człowieka w kontekście sztucznej inteligencji – implementacja tych zasad w skali krajowej i ich promocja na arenie międzynarodowej leży w interesie naszego państwa.

5. Edukacja społeczna i uodparnianie społeczeństwa. Nawet najlepsze systemy obronne mogą nie zapobiec wszystkim zagrożeniom, jeśli obywatele pozostaną nieświadomi i podatni na manipulacje. Dlatego kluczowym uzupełnieniem działań państwa musi być podniesienie kompetencji cyfrowych i medialnych społeczeństwa. Rekomenduje się włączenie tematyki

[311] European Commission, Tackling Online Disinformation: a European Approach, Komunikat Komisji COM(2018)236, 2018.

rozpoznawania fałszywych treści do programów nauczania – już na etapie szkoły średniej młodzież powinna uczyć się, jak krytycznie oceniać wiarygodność materiałów w Internecie i jak działają współczesne techniki manipulacji[312]. Organizacje państwowe mogą we współpracy z mediami publicznymi prowadzić kampanie informacyjne pokazujące przykłady *deepfake* (zarówno te szkodliwe, jak i nieszkodliwe – np. w filmach rozrywkowych), by obywatele potrafili je rozpoznać i nie dali się nabrać. Niezbędna jest także budowa zaufania poprzez transparentną komunikację władz. Rząd powinien jasno informować społeczeństwo o pojawiających się zagrożeniach informacyjnych – tak jak to miało miejsce na Ukrainie, gdzie oficjalnie ostrzegano przed możliwymi fałszywymi przemówieniami prezydenta, dzięki czemu, gdy takie wideo się pojawiło, wielu odbiorców od razu powątpiewało w jego prawdziwość. Władze muszą reagować szybko na głośne przypadki manipulacji – natychmiastowe zdementowanie plotki czy fałszywki przez wiarygodne źródła, najlepiej z pokazaniem dowodów na mistyfikację, może zniwelować dużą część jej szkodliwego efektu. Dlatego w ramach struktur państwa warto powołać zespół szybkiego reagowania komunikacyjnego, który w sytuacji kryzysu informacyjnego będzie odpowiedzialny za przekazanie społeczeństwu skoordynowanego, oficjalnego stanowiska wraz z analizą ekspercką. Taka rzetelna metakomunikacja wzmocni odporność społeczną, minimalizując panikę i dezinformacyjny chaos. Podsumowując, polityka bezpieczeństwa państwa w obliczu rozwoju analiz automatycznych musi opierać się na zasadzie wielowarstwowej obrony i równowagi. Jak wskazują eksperci, potrzebne jest jednoczesne działanie na wielu polach: technologicznym, prawnym, edukacyjnym i międzynarodowym[313]. Państwo powinno inwestować w nowoczesne narzędzia i kadry, ale równocześnie stworzyć gwarancje, że narzędzia te nie będą nadużywane przeciw obywatelom. Współpraca ze społeczeństwem – traktowanie obywateli jako sojuszników w walce z dezinformacją, a nie tylko jako „widzów" – jest kluczowa dla skuteczności strategii[314]. Tylko holistyczne podejście, łączące możliwości AI z mądrością ludzką i odpowiedzialną polityką, zapewni, że analiza treści

[312] Zob. C. Vaccari, A. Chadwick, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News, 6(1), 1-13.

[313] Zob. I. Kalpokas, J. Kalpokienė, Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation, Springer, 2022.

[314] Zob. C. Vaccari, A. Chadwick, Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News, 6(1), 1-13.

audiowizualnych stanie się trwałym wsparciem systemu bezpieczeństwa państwa, wzmacniając go, a nie destabilizując.

9.2. Rekomendacje dla instytucji bezpieczeństwa i mediów publicznych

W warunkach postępującej wizualizacji przekazu, rosnącej roli treści cyfrowych oraz gwałtownego przyrostu dezinformacji i manipulacji medialnych, kluczową rolę w przeciwdziałaniu zagrożeniom informacyjnym odgrywają instytucje odpowiedzialne za bezpieczeństwo państwa oraz media publiczne. Z uwagi na skalę i złożoność wyzwań, skuteczność działań w tym obszarze zależy od odpowiedniego przygotowania strukturalnego, kompetencyjnego oraz koordynacyjnego. Poniższe rekomendacje formułowane są z myślą o potrzebie strategicznej modernizacji i integracji systemu bezpieczeństwa informacyjnego Rzeczypospolitej Polskiej.

W pierwszej kolejności należy wskazać na potrzebę wzmocnienia zdolności analitycznych instytucji bezpieczeństwa, w tym służb wywiadowczych, kontrwywiadowczych oraz podmiotów cywilnych odpowiedzialnych za reagowanie kryzysowe. Dotyczy to zwłaszcza inwestycji w technologie rozpoznawania i analizy treści audiowizualnych, rozwijania narzędzi wykrywania manipulacji w materiałach cyfrowych (deepfake, syntetyczne audio, zmanipulowane obrazy), a także tworzenia wyspecjalizowanych zespołów zajmujących się monitoringiem i oceną przekazów medialnych w czasie rzeczywistym. Rekomenduje się utworzenie interdyscyplinarnych jednostek łączących kompetencje z zakresu technologii, bezpieczeństwa, komunikacji społecznej i psychologii poznawczej.

Kolejnym istotnym kierunkiem jest zacieśnienie współpracy pomiędzy instytucjami bezpieczeństwa a mediami publicznymi, w celu wypracowania spójnej polityki informacyjnej w sytuacjach kryzysowych. Media publiczne powinny pełnić rolę filaru wiarygodnej komunikacji, zdolnego do szybkiego reagowania na pojawiające się kampanie dezinformacyjne, korygowania fałszywych przekazów oraz edukowania społeczeństwa w zakresie rozpoznawania manipulacji. W tym kontekście kluczowe jest zapewnienie mediom publicznym niezależności redakcyjnej oraz oparcia ich działalności na standardach dziennikarstwa informacyjnego, a nie propagandowego.

Z punktu widzenia efektywnej komunikacji państwa z obywatelami rekomenduje się również wdrożenie strategii komunikacji kryzysowej, opartej na zasadzie przejrzystości, terminowości i dostępności informacji. Organy administracji publicznej – w tym Rządowe Centrum Bezpieczeństwa, Ministerstwo Spraw Wewnętrznych i Administracji, Ministerstwo Cyfryzacji – powinny posiadać z góry przygotowane, zaktualizowane procedury działania informacyjnego w różnych scenariuszach zagrożeń: od katastrof naturalnych, przez ataki cybernetyczne, po kampanie dezinformacyjne inspirowane zewnętrznie. Równie istotne jest testowanie tych procedur podczas ćwiczeń międzyresortowych i scenariuszy symulacyjnych.

Rekomenduje się ponadto włączenie mediów publicznych w długofalową strategię edukacji medialnej i cyfrowej. Ich zasięg oraz autorytet instytucjonalny powinny być wykorzystane do budowania odporności informacyjnej obywateli – szczególnie w zakresie rozpoznawania treści zmanipulowanych, zrozumienia działania algorytmów i platform cyfrowych oraz rozwijania umiejętności krytycznego odbioru komunikatów wizualnych. Można to realizować poprzez kampanie społeczne, programy edukacyjne, cykle tematyczne czy współpracę z ekspertami i organizacjami pozarządowymi.

Zaleca się również wprowadzenie systemu wczesnego ostrzegania o zagrożeniach dezinformacyjnych, który byłby zintegrowany z infrastrukturą mediów publicznych. Mechanizm taki mógłby pełnić funkcję analogiczną do Alertu RCB, ale w odniesieniu do przestrzeni informacyjnej – ostrzegając społeczeństwo przed groźnymi kampaniami manipulacyjnymi, próbami destabilizacji debaty publicznej czy zmasowanym rozpowszechnianiem fałszywych informacji. Skuteczność takiego systemu zależy od współdziałania państwa, mediów, sektora IT oraz organizacji monitorujących media społecznościowe.

Odrębną rekomendacją jest konieczność budowania zaufania społecznego do oficjalnych źródeł informacji. Nie wystarczy jedynie przeciwdziałać dezinformacji – państwo musi też być wiarygodnym i sprawnym nadawcą komunikatów. W tym celu należy prowadzić regularne działania komunikacyjne – także poza sytuacjami nadzwyczajnymi – pokazujące, że instytucje publiczne są kompetentne, otwarte i bliskie obywatelom. Jednym z elementów może być wzmocnienie obecności państwowych instytucji w mediach cyfrowych (w tym platformach wideo) z wykorzystaniem atrakcyjnych i przejrzystych form przekazu.

Ostatecznie, istotnym wyzwaniem jest systemowe uregulowanie relacji państwa z operatorami platform cyfrowych. Rekomenduje się aktywne zaangażowanie Polski w europejskie i międzynarodowe procesy regulacyjne dotyczące dezinformacji i sztucznej inteligencji, a także opracowanie krajowej polityki przeciwdziałania treściom zmanipulowanym, z jasno określonymi zasadami odpowiedzialności, współpracy i przejrzystości.

9.3. Edukacja medialna i cyfrowa jako narzędzie wzmacniania odporności społecznej

Każda ewolucja medialna niesie ze sobą istotne zmiany nie tylko na poziomie technologii, ale również praktyk społecznych[315].

Współczesna przestrzeń informacyjna cechuje się ogromnym natężeniem przekazów, ich wysoką dynamiką oraz trudnością w jednoznacznej ocenie wiarygodności. W takich warunkach jednym z kluczowych narzędzi wzmacniania odporności społecznej na zagrożenia dezinformacyjne i manipulacje medialne jest powszechna edukacja medialna i cyfrowa. Obejmuje ona zestaw kompetencji niezbędnych do świadomego, krytycznego i odpowiedzialnego uczestnictwa w życiu informacyjnym – zarówno jako odbiorca, jak i twórca treści. Jej znaczenie rośnie proporcjonalnie do skali wyzwań, jakie niesie za sobą funkcjonowanie w świecie zdominowanym przez obraz, szybki przekaz, sztuczną inteligencję i zautomatyzowane systemy rekomendacyjne.

Edukacja medialna i cyfrowa pełni dziś funkcję nie tylko kulturową i obywatelską, ale coraz częściej również bezpieczeństwa publicznego. To właśnie od poziomu świadomości informacyjnej obywateli zależy odporność społeczeństwa na próby destabilizacji przestrzeni publicznej poprzez fake newsy, deepfake'i, operacje wpływu czy zorganizowane kampanie dezinformacyjne. Z czasem, wraz z postępującym procesem profesjonalizacji i mediatyzacji komunikacji, w sferze online – osoby świadomie i intencjonalnie wpływające na innych, również z „czyjegoś nadania". W tej grupie

[315] G. Ptaszek, Edukacja medialna 3.0. Krytyczne rozumienie mediów cyfrowych w dobie Big Data i algorytmizacji, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2019, s. 21.

– celebryci i influenserzy, ale także podmioty zbiorowe, grupy interesu oraz –w XXI wieku – boty[316].

Podstawowym zadaniem edukacji w tym zakresie jest dostarczenie jednostkom narzędzi umożliwiających rozpoznawanie manipulacji wizualnej, ocenę intencji nadawcy komunikatu, odróżnianie faktów od opinii oraz zrozumienie procesów, jakie zachodzą w środowisku cyfrowym – od algorytmizacji przekazu po inżynierię danych.

Rola systemu edukacji formalnej w budowaniu tej odporności jest nie do przecenienia. Konieczne jest wprowadzenie do programów nauczania na wszystkich poziomach – od szkoły podstawowej po uniwersytet – elementów kształcenia w zakresie rozumienia przekazów medialnych i mechanizmów ich rozpowszechniania. Współczesny alfabet informacyjny nie ogranicza się już tylko do umiejętności czytania i pisania, ale obejmuje także kompetencje interpretacyjne, wizualne i cyfrowe, takie jak analiza materiału wideo, ocena wiarygodności źródła, rozpoznawanie zmanipulowanych grafik, a nawet podstawowa wiedza o działaniu platform społecznościowych, botów informacyjnych czy generatywnej sztucznej inteligencji.

Oprócz edukacji szkolnej, istotną rolę w przeciwdziałaniu dezinformacji odgrywają kampanie społeczne i programy informacyjne realizowane przez instytucje publiczne, organizacje pozarządowe oraz media. Ich celem jest dotarcie do szerokich grup społecznych z przekazem edukacyjnym, który w atrakcyjny, dostępny i prosty sposób uświadamia istnienie zagrożeń i uczy podstawowych strategii obrony przed manipulacją. Kampanie takie powinny być realizowane w sposób systematyczny, z wykorzystaniem nowoczesnych form przekazu (wideo, grafiki, animacje interaktywne), a także dostosowane do różnych grup wiekowych i poziomów kompetencji technologicznych. Działania tego typu mogą stanowić skuteczną odpowiedź na dynamiczne zmiany w środowisku medialnym, zwłaszcza gdy są prowadzone w partnerstwie z mediami publicznymi i edukacyjnymi. W niniejszym zestawieniu przedstawiono porównanie kompetencji kluczowych w ujęciu z lat 2015 i 2025. Tabela przedstawia kierunki zmian w podejściu do kompetencji, które są uznawane za niezbędne dla rozwoju jednostki oraz funkcjonowania w dynamicznie zmieniającym się społeczeństwie informacyjnym i edukacji.

[316] A. Ogonowska, Od intrygi i spisku do zagrożeń hybrydowych. Nowe kompetencje i edukacja medialna w cyberkulturze 3.0, Studia de Cultura, 14(1), 7.

Tabela 8. Porównanie kompetencji kluczowych w 2015 i 2025 roku

Kompetencje kluczowe w 2015 roku	Kompetencje kluczowe w 2025 roku (prognozowane/postulowane)
Porozumiewanie się w języku ojczystym	Kompetencje komunikacyjne i retoryczne (w mowie i piśmie, w tym w debacie publicznej)
Porozumiewanie się w językach obcych	Kompetencje interkulturowe i wielojęzyczność
Kompetencje matematyczne oraz naukowo-techniczne	Kompetencje analityczne i rozwiązywania problemów w środowisku STEAM
Kompetencje cyfrowe	Kompetencje cyfrowe i w zakresie sztucznej inteligencji (AI literacy)
Umiejętność uczenia się	Samodzielne uczenie się i zarządzanie wiedzą w środowisku informacyjnym
Kompetencje osobiste, społeczne i w zakresie uczenia się	Kompetencje społeczno-emocjonalne, rezyliencja i inteligencja emocjonalna
Kompetencje w zakresie przedsiębiorczości	Kompetencje innowacyjne i przedsiębiorczość społeczna
Kompetencje w zakresie świadomości i ekspresji kulturalnej	Kompetencje kulturowe, estetyczne i medialne
-	Kompetencje w zakresie zrównoważonego rozwoju i odpowiedzialności społecznej
-	Kompetencje w zakresie bezpieczeństwa cyfrowego i informacyjnego

Źródło: opracowanie własne.

Analiza porównawcza kompetencji kluczowych pokazuje wyraźne przesunięcie akcentów z umiejętności podstawowych na kompetencje złożone, oparte na myśleniu krytycznym, umiejętności rozwiązywania problemów, a także kompetencje społeczne, cyfrowe i ekologiczne. Zwiększa się znaczenie wielowymiarowej gotowości do uczenia się przez całe życie, świadomości kulturowej, obywatelskiej oraz adaptacji do wyzwań związanych z rozwojem technologii i zmianami cywilizacyjnymi.

Ważnym elementem edukacji medialnej i cyfrowej są również programy lokalne i obywatelskie, które wspierają działania oddolne, organizowane przez biblioteki, domy kultury, uniwersytety trzeciego wieku, koła gospodyń wiejskich czy lokalne organizacje pozarządowe. To właśnie w przestrzeni lokalnej możliwe jest bezpośrednie kształtowanie nawyków informacyjnych, organizowanie warsztatów, prowadzenie zajęć dla seniorów czy budowanie mostów pokoleniowych w zakresie kompetencji cyfrowych. Takie inicjatywy mogą znacząco podnieść poziom odporności informacyjnej społeczności lokalnych, które są szczególnie podatne na dezinformację – zwłaszcza w warunkach ograniczonego dostępu do niezależnych źródeł informacji czy edukacji.

Jednakże rozwój edukacji medialnej i cyfrowej w Polsce – podobnie jak w wielu innych państwach – napotyka liczne bariery[317]. Do najważniejszych należą brak systemowych rozwiązań w polityce oświatowej, niedostateczne przygotowanie nauczycieli i edukatorów, niedobór nowoczesnych materiałów dydaktycznych oraz wciąż niska świadomość społeczna co do znaczenia tej formy edukacji. Media społecznościowe stały się elementem uwarunkowania bytowego: środkiem komunikacji oraz „miejscem" pracy cyfrowej[318]. Potrzebna jest zatem polityka publiczna, która traktuje edukację medialną nie jako fakultatywny dodatek, lecz jako strategiczny komponent bezpieczeństwa państwa i integralny element kształcenia obywatelskiego.

W perspektywie długofalowej to właśnie edukacja – zarówno formalna, jak i nieformalna – ma największy potencjał w zakresie wzmacniania odporności społecznej. Dzięki niej jednostki uczą się nie tylko jak przetrwać w natłoku informacji, ale także jak aktywnie uczestniczyć w życiu społecznym, podejmować świadome decyzje, bronić wartości demokratycznych i współtworzyć bezpieczne środowisko informacyjne. Edukacja medialna i cyfrowa, odpowiednio zaprojektowana i wspierana przez instytucje publiczne oraz sektor prywatny, może stać się skuteczną tarczą ochronną wobec złożonych zagrożeń XXI wieku – w tym tych, które przybierają formę obrazów, dźwięków i narracji audiowizualnych.

9.4. Możliwości dalszych badań nad dezinformacją audiowizualną

Biorąc pod uwagę powyższe wnioski, zidentyfikowano szereg obszarów wymagających pogłębionych badań i rozwoju w kontekście dezinformacji audiowizualnej:

- Nowe algorytmy detekcji o wyższej odporności – Konieczne jest opracowanie kolejnej generacji metod wykrywania deepfake, charakteryzujących się wyższą skutecznością oraz zdolnością do generalizacji

[317] N. Pawłowna Niedzwiedskaja, Społeczeństwo cyfrowe: problemy i perspektywy rozwoju [w:] red. J. M. Osipow, A. Z. Nowak, Rewolucja cyfrowa. Wyzwania, problemy, perspektywy rozwoju, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa 2019, s. 16.

[318] J. Kozubowski, Socjologia wiedzy społeczeństwa cyfrowego – „praca cyfrowa" a ideologia mediów społecznościowych, „Nierówności Społeczne a Wzrost Gospodarczy", 66, 76.

na nieznane typy fałszerstw. Istotne jest, aby proces detekcji uwzględniał nie tylko analizę cech wizualnych, lecz także kontekst i znaczenie przekazu. Rozwój adaptacyjnych, kontekstowych algorytmów, zdolnych między innymi do wykrywania niespójności między obrazem a towarzyszącym mu opisem sytuacji, stanowi obiecujący kierunek dalszych badań.

- Analiza multimodalna – Jednym z najważniejszych kierunków rozwoju jest integracja analiz różnych modalności (obrazu, dźwięku, tekstu) w celu wychwytywania manipulacji. Obecne prace pokazują, że modele jednocześnie badające warstwę wizualną nagrania oraz ścieżkę audio osiągają lepsze wyniki od klasycznych metod jednokanałowych. Wykorzystanie niespójności między ruchami ust a głosem, między intonacją a treścią wypowiedzi czy między obrazem a generowanymi odgłosami pozwala wykryć fałszerstwa, które umykają pojedynczym detektorom. Przyszłe badania powinny zatem koncentrować się na metodach łączonych (tzw. multimodalnych), zdolnych do kompleksowej analizy materiału wideo wraz z dźwiękiem, a być może także z towarzyszącym tekstem (np. automatycznie rozpoznawanymi wypowiedziami czy opisami). Tego rodzaju podejście zwiększa wiarygodność oceny autentyczności przekazu, co potwierdzają już pierwsze eksperymenty naukowe.
- Uwierzytelnianie treści za pomocą *blockchain* – Obszar ten stanowi interesujący kierunek badań, koncentrujący się na zastosowaniu technologii *blockchain* do weryfikacji autentyczności materiałów audiowizualnych. Jako niezmienny, rozproszony rejestr transakcji, *blockchain* może pełnić funkcję globalnej bazy dowodowej dla plików multimedialnych. Każdy oryginalny film lub nagranie dźwiękowe mogłoby być opatrywane w momencie powstania kryptograficznym podpisem oraz znacznikiem czasu, zapisanym w *blockchain*, co praktycznie uniemożliwiałoby przeprowadzenie manipulacji bez wykrycia. Badania w tym obszarze powinny obejmować opracowanie protokołów umożliwiających weryfikację autentyczności źródła, na przykład poprzez sprawdzanie, czy dany plik posiada ważny podpis cyfrowy zaufanej instytucji lub producenta sprzętu. Obecnie pojawiają się prototypowe rozwiązania oparte na tej technologii, stosowane w branży medialnej do znakowania treści. Konieczne jest zbadanie ich skuteczności oraz

możliwości skalowania, a także potencjału *blockchain* do umożliwienia zdecentralizowanej weryfikacji poprzez konsensus wielu uczestników sieci. Choć sama technologia *blockchain* nie eliminuje problemu dezinformacji, może stać się istotnym elementem systemu zwiększającego zaufanie do informacji cyfrowych.

- Detekcja nowych form treści syntetycznych – Postępujący rozwój generatywnej sztucznej inteligencji powoduje, że oprócz klasycznych *deepfake'ów* wideo coraz częściej pojawiają się inne rodzaje syntetycznych mediów, takie jak imitacje głosu, wirtualni prezenterzy, fałszywe obrazy (np. zdjęcia tworzone przez sieci neuronowe) czy syntetyczne dane sensoryczne, w tym podrabiane nagrania radarowe lub satelitarne. Wobec tego konieczne jest, aby dalsze badania nad detekcją manipulacji cyfrowych nadążały za tym trendem, opracowując metody rozpoznawania fałszywych treści we wszystkich tych domenach. Badania powinny skoncentrować się na opracowaniu algorytmów zdolnych do wykrywania syntetycznej mowy, na przykład poprzez analizę nienaturalnych intonacji lub specyficznych artefaktów powstających w procesie generacji dźwięku, jak również algorytmów wykrywających generowane obrazy, poprzez identyfikację anomalii w rozkładzie szumów, czy nieregularności w oświetleniu i cieniach. Przyszłe rozwiązania powinny również kłaść nacisk na analizę semantyczną i logiczną przekazu, na przykład poprzez sprawdzanie zgodności prezentowanych faktów z wiarygodnymi źródłami, co stanowi wyzwanie interdyscyplinarne na pograniczu sztucznej inteligencji i nauk o informacji. Uzupełnieniem tych metod może być rozwój technik znakowania treści oryginalnych, takich jak cyfrowe „odciski palca" kamer i mikrofonów, co ułatwi odróżnienie autentycznych materiałów od syntetycznych imitacji. W rezultacie, integracja metod detekcji post factum z mechanizmami weryfikacji u źródła może stworzyć synergiczny, bardziej skuteczny system zabezpieczeń, zdolny do kompleksowego przeciwdziałania manipulacjom treści cyfrowych.

Odporność społeczna i edukacja odbiorców – Techniczne rozwiązania muszą być wspierane przez dogłębne zrozumienie czynników społecznych, które determinują skuteczność dezinformacji. Konieczne są badania mające na celu opracowanie metod zwiększających odporność społeczeństwa na fałszywe przekazy, m.in. poprzez systematyczną edukację medialną, kampanie

informacyjne oraz wdrażanie strategii pre-*bunkingu* (uprzedzającego ostrzegania przed manipulacją). Literatura przedmiotu podkreśla, że podniesienie kompetencji medialnych obywateli znacząco utrudnia rozprzestrzenianie się *fake newsów*, gdyż osoby wykształcone w zakresie krytycznego odbioru treści wykazują mniejszą podatność na dezinformację oraz rzadziej udostępniają zmanipulowane materiały. Badania eksperymentalne wykazały, że nawet krótkie interwencje edukacyjne mogą obniżać skłonność użytkowników do akceptowania i rozpowszechniania *deepfake'ów*. W związku z powyższym, dalsze prace naukowe powinny koncentrować się na rozwijaniu inicjatyw popularyzujących wiedzę o mechanizmach dezinformacji, na przykład poprzez integrację tych zagadnień w programach nauczania oraz tworzenie dostępnych kursów online, dedykowanych różnym grupom wiekowym. Istotnym kierunkiem badań jest również analiza psychologicznych aspektów odbioru fałszywych treści – w tym identyfikacja czynników, które wpływają na akceptację dezinformacji, mechanizmów weryfikacji informacji przez odbiorców oraz skutecznych strategii korygowania fałszywych przekonań. Interdyscyplinarne badania, łączące perspektywy nauk o bezpieczeństwie, psychologii, socjologii oraz nauk o komunikacji, są kluczowe dla opracowania kompleksowych metod budowania odporności społecznej na dezinformację.

9.5. Strategie walki z *deepfake* i manipulacją medialną na poziomie państwowym

Rosnące zaawansowanie technik generowania *deepfake'ów* oraz innych syntetycznych obrazów i dźwięków stawia przed badaczami i instytucjami nowe wyzwania w zakresie ich detekcji. Dotychczasowe metody oparte na wyszukiwaniu artefaktów (np. nienaturalne piksele czy zaburzenia kolorów) okazują się coraz mniej skuteczne w warunkach rzeczywistych, ponieważ twórcy fałszywek doskonalą algorytmy generatywne, minimalizując widoczne ślady manipulacji. W konsekwencji trwa „wyścig zbrojeń" między technologiami tworzenia a technologiami wykrywania – każda nowa metoda detekcji skłania fałszerzy do opracowania kontrstrategii omijających zabezpieczenia. Już dziś badacze wskazują, że poleganie wyłącznie na analizie cech technicznych materiału (takich jak np. nienaturalne ruchy ust czy

artefakty kompresji) może nie wystarczyć, gdyż przyszłe generatory będą potrafiły je imitować lub maskować. W odpowiedzi na te wyzwania zarysowuje się kilka nowych kierunków badań i działań technologicznych mających na celu wzmocnienie skuteczności wykrywania fałszerstw audiowizualnych. Jednym z obiecujących podejść jest analiza multimodalna, łącząca informacje z różnych źródeł i modalności. Oznacza to, że zamiast badać osobno obraz, dźwięk czy tekst, systemy detekcji porównują spójność między tymi warstwami przekazu. Na przykład głos w materiale wideo może być weryfikowany pod kątem zgodności z ruchami ust postaci lub z wcześniej zarejestrowanymi wypowiedziami danej osoby. Dzięki uczeniu maszynowemu możliwe staje się wykrywanie subtelnych niespójności między obrazem a dźwiękiem, które dla ludzkiego oka i ucha pozostają niewidoczne. Badania wskazują, że sieci neuronowe mogą identyfikować anomalie w sygnałach audio po przekształceniu ich na spektrogram (wizualizację dźwięku), traktując wykrywanie fałszywego głosu jak problem rozpoznawania obrazu. Równocześnie algorytmy analizujące obraz wideo są trenowane do wychwytywania mikroruchów i detali fizjologicznych – przykładowo nieregularnego mrugania, nienaturalnej mimiki czy niespójnych kierunków patrzenia – które mogą świadczyć o podmianie twarzy. Takie holistyczne, wielomodalne podejście pozwala na skuteczniejszą detekcję, ponieważ fałszerstwo musi oszukać jednocześnie wiele czujników: zarówno analizę wizualną, jak i dźwiękową czy tekstową. Kolejnym rozwijanym kierunkiem jest wykorzystanie sztucznej inteligencji nie tylko do generowania, ale i do wykrywania fałszywych treści. Paradoksalnie, algorytmy oparte na podobnych technikach, co generatory *deepfake*, mogą służyć do ich identyfikacji. Sieci neuronowe uczą się rozpoznawać charakterystyczne dla syntetycznych mediów wzorce – tzw. „artefakty generatywne" – na podstawie obszernych baz prawdziwych i fałszywych nagrań. Przykładowo, modele detekcyjne potrafią wykryć nienaturalne rozmycia lub pikselizacje w określonych klatkach wideo, zaburzenia oświetlenia na krawędziach twarzy czy nielogiczne cienie, które są efektem działania algorytmu *deepfake*. Sztuczna inteligencja może też monitorować przepływ informacji online: systemy wyposażone w mechanizmy uczenia głębokiego skanują media społecznościowe i serwisy internetowe w poszukiwaniu materiałów o cechach wskazujących na manipulację. Już teraz w ramach międzynarodowych inicjatyw powstają narzędzia automatycznie śledzące pojawianie się podejrzanych treści

– przykładem jest platforma Hamilton 2.0, która w czasie rzeczywistym monitoruje i analizuje przekazy propagandowe w mediach społecznościowych. Również NATO inwestuje w analitykę danych opartą na AI dla celów wczesnego ostrzegania – rozwinięto system zdolny do przetwarzania ogromnych wolumenów danych medialnych i wychwytywania w nich sygnałów świadczących o skoordynowanych kampaniach dezinformacyjnych. Takie rozwiązania przyspieszają reakcję na pojawiające się *deepfake'i*, pozwalając organom państwowym lub firmom na szybkie oznaczenie danego nagrania jako podejrzane i skierowanie go do weryfikacji eksperckiej, zanim zdąży się szeroko rozpowszechnić. Niemniej, wykorzystanie AI w detekcji niesie też wyzwania – modele uczą się na podstawie dostępnych danych i mogą mieć trudność z rozpoznaniem zupełnie nowych, niewidzianych wcześniej typów fałszerstw. Dlatego kluczowe jest ciągłe doskonalenie i trening tych systemów na aktualizowanych zbiorach *deepfake'ów*, a także łączenie ich z innymi metodami dla zwiększenia niezawodności. Coraz większą uwagę poświęca się również weryfikacji kontekstowej oraz uwierzytelnianiu u źródła jako dopełnieniu czysto technicznej detekcji. Pierwszy z tych elementów polega na analizie otoczenia informacji: sprawdzaniu, czy przedstawione w materiale zdarzenie miało miejsce w rzeczywistości (np. poprzez konfrontację z oficjalnymi komunikatami lub harmonogramem wystąpień danej osoby), śledzeniu drogi, jaką dana treść przebyła w mediach społecznościowych, oraz identyfikacji ewentualnych powiązań z wcześniej znanymi kampaniami dezinformacyjnymi. Tego rodzaju analiza kontekstu pozwala wykryć fałsz nawet wtedy, gdy sam materiał audiowizualny jest technicznie trudny do zakwestionowania – np. film przedstawiający przemówienie polityka może wyglądać autentycznie, ale zostanie zdemaskowany jako *deepfake*, jeśli polityk ten w rzeczywistości nigdy takiego przemówienia nie wygłosił ani nie przebywał w miejscu pokazanym na nagraniu. Drugi element – uwierzytelnianie u źródła – polega na potwierdzaniu autentyczności materiałów już w momencie ich powstania i dystrybucji. W praktyce sprowadza się to do znakowania treści oryginalnych specjalnymi znacznikami cyfrowymi, które pozwolą w przyszłości zweryfikować, czy plik został zmodyfikowany. Cyfrowe podpisy i znaki wodne mogą być osadzone w metadanych zdjęcia czy filmu przez autentycznych twórców (np. przez agencje prasowe, instytucje rządowe, redakcje), co utrudni ingerencję bez pozostawienia śladu. W ostatnich latach branża technologiczna zawiązała koalicje na rzecz

standaryzacji takich rozwiązań – przykładem jest inicjatywa Project Origin wspierana m.in. przez BBC, Microsoft i Adobe, dążąca do stworzenia otwartego standardu proweniencji mediów. Koncepcja ta zakłada, że każdy opublikowany materiał mógłby otrzymać unikalny „odcisk palca" (*hash* lub inny znacznik) zapisany w rozproszonej bazie danych o wysokim poziomie zabezpieczeń kryptograficznych (analogicznie do technologii *blockchain*). Następnie przeglądarki internetowe lub aplikacje mogłyby automatycznie porównać materiał, na który natrafia użytkownik, z oryginalnym wpisem w takiej bazie – i ostrzec, jeśli treść nie będzie zgodna. Tego typu systemy wiarygodności treści pozwoliłyby na błyskawiczne wychwycenie prób manipulacji (np. podmiany fragmentu obrazu lub dźwięku) i tym samym stanowiłyby fundament infrastruktury zaufania w sferze informacji cyfrowej. Warto podkreślić, że wdrożenie uwierzytelniania u źródła wymaga szerokiej współpracy – ustanowienia standardów akceptowanych zarówno przez dostawców treści, platformy internetowe, jak i organy państwowe. Już obecnie trwają prace normalizacyjne w tym zakresie, również na forum międzynarodowym, co wskazuje, iż podejście to może stać się w przyszłości istotnym filarem przeciwdziałania fałszerstwom audiowizualnym. Podsumowując, walka z *deepfake'ami* wymaga wielowątkowych innowacji: od coraz bystrzejszych algorytmów AI zdolnych do analizy treści w wielu wymiarach jednocześnie, przez systemy znakowania i śledzenia pochodzenia materiałów, po inteligentne mechanizmy włączające kontekst i wiedzę o świecie do procesu weryfikacji. Żaden z tych elementów osobno nie gwarantuje pełnego sukcesu – ich skuteczność rośnie, gdy są stosowane komplementarnie. Co więcej, eksperci zgodnie zaznaczają, że sama detekcja to nie wszystko: nawet idealne wykrycie fałszywki nie zapobiegnie szkodom, jeśli nastąpi zbyt późno albo jeśli społeczeństwo zignoruje sprostowanie. Dlatego nowe technologie powinny iść w parze z budowaniem mechanizmów szybkiej reakcji (np. protokołów alarmowych i kampanii informacyjnych korygujących narrację) oraz z działaniami edukacyjnymi zwiększającymi świadomość istnienia takich zagrożeń.

9.6. Wnioski

Przeprowadzone analizy jednoznacznie wskazują, że dezinformacja audiowizualna stanowi jedno z najpoważniejszych zagrożeń dla bezpieczeństwa państwa oraz integralności społecznej. W kontekście dynamicznego rozwoju mediów cyfrowych, fałszywe przekazy w formie nagrań wideo i audio nie tylko podważają zaufanie do instytucji publicznych, ale również mają potencjał destabilizacji procesów demokratycznych. Jak podkreśla Światowe Forum Ekonomiczne, dezinformacja została uznana za jedno z kluczowych globalnych ryzyk roku 2024, a *deepfake* – jako jedna z najbardziej niepokojących technologii wykorzystywanych w generowaniu fałszywych treści[319].

Badania wykazały, iż choć nowoczesne technologie oparte na sztucznej inteligencji umożliwiają automatyczne monitorowanie i analizę treści audiowizualnych, to jednak istniejące systemy detekcji nadal borykają się z istotnymi ograniczeniami. Zaawansowane algorytmy, choć zdolne do identyfikacji subtelnych anomalii biologicznych czy technicznych, osiągają w warunkach rzeczywistych skuteczność rzędu około 65% w rozpoznawaniu dotychczas nieznanych *deepfake'ów*[320]. W związku z tym samo wykrycie manipulacji nie gwarantuje skutecznego zapobieżenia rozprzestrzenianiu się dezinformacji – fałszywe treści mogą zdobyć ogromny zasięg, zanim zostaną publicznie zdementowane lub oznaczone jako nieprawdziwe. Analizy podkreślają również dwojaki charakter roli sztucznej inteligencji – z jednej strony technologia ta umożliwia tworzenie hiperrealistycznych fałszerstw, z drugiej zaś stanowi istotne narzędzie w walce z dezinformacją. Integracja algorytmów AI w systemach monitorowania przestrzeni informacyjnej przynosi wymierne korzyści, umożliwiając szybką identyfikację manipulacji nawet w transmisjach na żywo. Przykładowo, system *Information Environment Assessment* wdrożony przez NATO w 2024 roku pozwala na analizę ogromnych zbiorów danych w czasie rzeczywistym, co umożliwia wczesne wykrycie wrogich narracji[321]. Jednakże, niezależnie od postępów technologicznych, wyniki badań wskazują, że poleganie wyłącznie na

[319] Światowe Forum Ekonomiczne, Global Risks Report, 2024.

[320] U.S. Government Accountability Office (GAO), Raport Assessment of Deepfake Detection Technologies, Waszyngton, 2023.

[321] NATO, Information Environment Assessment System Implementation Report, 2024.

narzędziach technicznych jest niewystarczające. Konieczna jest holistyczna strategia łącząca zaawansowane technologie, odpowiednie regulacje prawne oraz działania edukacyjne w celu budowania odporności społeczeństwa na dezinformację. Ważnym wnioskiem wynikającym z przeprowadzonych badań jest konieczność podejścia interdyscyplinarnego. Skuteczne przeciwdziałanie zagrożeniom wynikającym z dezinformacji audiowizualnej wymaga nie tylko udoskonalenia technologii detekcji, ale także wdrożenia kompleksowych rozwiązań w sferze legislacyjnej oraz systematycznej edukacji społecznej. Współczesne systemy bezpieczeństwa państwa muszą integrować mechanizmy wczesnego ostrzegania z efektywnymi procedurami reagowania, aby szybko neutralizować zagrożenia zanim osiągną one skalę krytyczną[322].

Podsumowując, badania dowodzą, że:

- Dezinformacja audiowizualna stanowi poważne zagrożenie dla stabilności państwa i procesów demokratycznych, szczególnie w kontekście rosnącej popularności *deepfake'ów*.
- Technologie analizy treści, oparte na sztucznej inteligencji, choć dynamicznie się rozwijają, nadal wymagają dalszych ulepszeń, zwłaszcza w warunkach rzeczywistych zastosowań.
- Sztuczna inteligencja pełni funkcję dwojaką – jako narzędzie tworzenia fałszywych treści, ale także jako kluczowy element systemów obronnych w walce z dezinformacją.
- Skuteczna strategia przeciwdziałania dezinformacji musi łączyć zaawansowane narzędzia technologiczne, adekwatne regulacje prawne oraz programy edukacji medialnej, co pozwoli na budowanie odporności społecznej na zagrożenia informacyjne.

Dalsze badania w tym obszarze powinny koncentrować się na wieloaspektowej optymalizacji algorytmów detekcyjnych, co obejmuje zarówno udoskonalenie technik wykrywania fałszerstw audiowizualnych, jak i opracowanie uniwersalnych protokołów walidacji. Aktualne modele oparte na głębokich sieciach neuronowych, choć efektywne w warunkach laboratoryjnych, wykazują ograniczenia przy zastosowaniu w realnych, dynamicznie zmieniających się środowiskach informacyjnych. W związku z tym konieczne jest przeprowadzenie badań empirycznych, które pozwolą na

[322] Komisja Europejska, Strengthened Code of Practice on Disinformation 2022, Bruksela: Komisja Europejska, 2022.

adaptację algorytmów do warunków wielomodalnych, gdzie obraz, dźwięk oraz metadane powinny być analizowane jednocześnie. W kontekście tych prac, kluczowe znaczenie mają badania porównawcze, które umożliwią identyfikację optymalnych parametrów oraz metod fuzji danych z różnych modalności, co przyczyni się do zwiększenia skuteczności detekcji *deepfake'ów* i innych manipulacji cyfrowych[323].

Równolegle z technologicznymi wyzwaniami, dalsze badania powinny również skupiać się na opracowaniu międzynarodowych standardów współpracy. Współpraca między państwami i instytucjami badawczymi jest niezbędna, aby zapewnić szybką wymianę informacji oraz koordynację działań w zakresie przeciwdziałania dezinformacji. Ujednolicenie procedur, wspólne platformy wymiany danych oraz stworzenie standardowych metryk oceny skuteczności systemów detekcyjnych mogłyby znacząco podnieść jakość monitoringu globalnej przestrzeni informacyjnej. Takie działania powinny obejmować zarówno organy państwowe, jak i niezależne instytucje badawcze oraz podmioty z sektora prywatnego, co umożliwi integrację różnorodnych doświadczeń i perspektyw naukowych. Ważnym aspektem dalszych badań jest także zbadanie implikacji etycznych oraz prawnych wynikających z masowej automatyzacji detekcji manipulacji. Wprowadzenie międzynarodowych standardów wymaga nie tylko technicznej interoperacyjności systemów, ale również zapewnienia, że mechanizmy te nie naruszają fundamentalnych praw obywatelskich, takich jak prywatność i wolność słowa. W związku z tym, interdyscyplinarne zespoły badawcze, łączące ekspertów z dziedzin prawa, etyki oraz informatyki, powinny pracować nad opracowaniem kompleksowych ram regulacyjnych, które uwzględnią zarówno potrzeby bezpieczeństwa państwa, jak i ochronę praw jednostki. Takie kompleksowe podejście przyczyni się do stworzenia efektywnych systemów obronnych w przestrzeni informacyjnej, zwiększając odporność państw na zagrożenia związane z dezinformacją i manipulacją audiowizualną

[323] U.S. Government Accountability Office (GAO), Raport Assessment of Deepfake Detection Technologies, Waszyngton, 2023.

Podsumowanie

Podsumowując przeprowadzone rozważania, niniejsza monografia kompleksowo przedstawia problematykę analizy treści audiowizualnych w kontekście współczesnych systemów bezpieczeństwa państwa. Ukazuje ona interdyscyplinarny charakter zagadnienia, które obejmuje zarówno aspekty technologiczne, społeczne, psychologiczne, jak i prawne. W wyniku dynamicznego rozwoju technologii cyfrowych państwa znalazły się wobec konieczności stawienia czoła nowemu rodzajowi zagrożeń – manipulacjom medialnym, szczególnie w formie treści audiowizualnych, które cechuje wysoki poziom oddziaływania na odbiorców. Analiza przedstawiona w monografii jednoznacznie wskazuje, że treści audiowizualne są niezwykle efektywnym narzędziem manipulacji informacyjnej, mogącym znacząco wpływać na społeczeństwa, procesy polityczne oraz bezpieczeństwo narodowe.

Szczególną uwagę poświęcono metodom analitycznym stosowanym do wykrywania manipulacji audiowizualnych, podkreślając kluczowe znaczenie wielomodalności – integracji analiz obrazu, dźwięku oraz tekstu. Zaprezentowane wyniki badań empirycznych potwierdzają wysoką efektywność metod bazujących na głębokim uczeniu oraz zaawansowanych technologiach cyfrowej analizy forensycznej, umożliwiających identyfikację nawet subtelnych form dezinformacji. Jednocześnie wskazano na istniejące ograniczenia technologiczne, które wynikają z ciągłego rozwoju algorytmów generatywnych stosowanych do tworzenia realistycznych falsyfikatów audiowizualnych, takich jak *deepfake*. Przeprowadzone analizy wskazały na konieczność nieustannego doskonalenia narzędzi detekcji oraz systematycznego aktualizowania algorytmów, aby zapewnić ich efektywność wobec coraz bardziej wyrafinowanych technik manipulacji. Studia przypadków omówione w monografii ukazały realny wymiar zagrożeń związanych z manipulacjami audiowizualnymi, które są wykorzystywane

do destabilizacji politycznej, społecznej i militarnej. Analiza tych przypadków pozwoliła na zidentyfikowanie psychologicznych i społecznych mechanizmów sprzyjających rozprzestrzenianiu dezinformacji, w tym efektu potwierdzenia, mechanizmów selektywnej ekspozycji oraz zjawiska tworzenia baniek informacyjnych. Szczególnie istotnym zjawiskiem jest efekt tzw. „dywidendy kłamcy", czyli sytuacji, w której sama świadomość możliwości generowania realistycznych fałszerstw prowadzi do erozji zaufania do wszelkich materiałów audiowizualnych, co z kolei sprzyja dalszemu rozpowszechnianiu chaosu informacyjnego.

Rekomendacje zawarte w pracy wskazują na konieczność wdrażania kompleksowych strategii przeciwdziałania zagrożeniom audiowizualnym, obejmujących jednoczesny rozwój technologiczny, edukację społeczną oraz spójne regulacje prawne. W szczególności zalecane jest szerokie wprowadzenie technologii blockchain oraz innych metod kryptograficznych do znakowania oryginalnych treści audiowizualnych, co znacząco utrudnia późniejsze manipulacje. Z kolei rozwiązania legislacyjne powinny równoważyć ochronę przestrzeni informacyjnej z poszanowaniem wolności słowa oraz prywatności obywateli. Z perspektywy dalszych badań kluczowe będzie pogłębianie współpracy między środowiskiem akademickim, instytucjami państwowymi oraz sektorem technologicznym w celu opracowywania nowych metod analitycznych, narzędzi detekcji oraz standardów prawnych. Ważnym kierunkiem rozwoju jest także integracja sztucznej inteligencji z technologiami analizy Big Data, co pozwoli na stworzenie systemów zdolnych do analizy ogromnych zbiorów danych audiowizualnych w czasie rzeczywistym. Ponadto, przyszłość należy do dalszego rozwoju metod uwierzytelniania treści oraz upowszechniania praktyk *pre-bunkingu* i edukacji medialnej społeczeństwa, co będzie istotnym elementem wzmacniania ogólnej odporności informacyjnej państwa. W konsekwencji przeprowadzone analizy potwierdzają, że kompleksowe, interdyscyplinarne podejście stanowi jedyną skuteczną strategię w walce z narastającymi zagrożeniami informacyjnymi. Niniejsza monografia stanowi znaczący wkład w nauki o bezpieczeństwie, medioznawstwo oraz informatykę, a także otwiera drogę do dalszych badań nad udoskonalaniem metod ochrony przestrzeni informacyjnej.

Literatura

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security (WIFS)*.

Agarwal, S., El-Gaaly, T., Farid, H., & Lim, S.-N. (2020). Detecting deep-fake videos from appearance and behavior. *arXiv preprint*. https://arxiv.org/abs/2004.14491

Akinyele, D., & Olaoye, G. (2024). Human-in-the-loop misinformation detection: Collaborative approach with human fact-checkers. *ResearchGate*.

Aleksandrowicz, T. R. (2018). Bezpieczeństwo informacyjne państwa. *Studia Politologiczne*, *49*, 33-50.

Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2011). A SIFT-based forensic method for copy–move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, *6*(3), 1099-1110. https://doi.org/10.1109/TIFS.2011.2129512

Amerini, I., Barni, M., Battiato, S., Bestagini, P., et al. (2024). Deepfake media forensics: State of the art and challenges ahead. *arXiv preprint*. https://doi.org/10.48550/arXiv.2408.00388

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154-175. https://doi.org/10.1080/21670811.2017.1345645

Balcerowicz, B. (2010). *Siły zbrojne w czasie pokoju, kryzysu i wojny*. Wydawnictwo Naukowe Scholar.

Batorowska, H. (2021). *Kultura bezpieczeństwa informacyjnego w środowisku walki o przewagę informacyjną*. Wydawnictwo Libron.

Batorowska, H., Motylińska, P., & Wasiuta, O. (2020). *Bezpieczeństwo informacyjne i medialne w czasach nadprodukcji informacji*. Wydawnictwo Naukowe i Edukacyjne SBP.

Bay, H., Tuytelaars, T., & Van Gool, L. (2008). SURF: Speeded up robust features. *Computer Vision and Image Understanding*, *110*(3), 346-359. https://doi.org/10.1016/j.cviu.2007.09.014

Bączek, P. (2005). *Zagrożenia informacyjne a bezpieczeństwo państwa polskiego*. Wydawnictwo Adam Marszałek.

Bąk, T., & Ciekanowski, Z. (2012). *Teorie bezpieczeństwa*. PWSTE.

Bąk, T. (2020). Rola mediów we współczesnych konfliktach zbrojnych i aktach terrorystycznych. WSIiZ.

Bhakt, R. (2019). WhatsApp rumours and lynchings: Mob violence in India. *Economic & Political Weekly*, *54*(6), 41-44.

Bielawski, R., Grenda, B., & Majdan, P. (2017). Wieloaspektowa ewaluacja wykorzystania mediów społecznościowych na potrzeby kierowania SBN RP (ewaluacji ryzyka cyberzagrożeń BN). Akademia Sztuki Wojennej.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Brangetto, P., & Veenendaal, M. A. (2016). Influence cyber operations: The use of cyberattacks in support of influence operations. In *International Conference on Cyber Conflict (CyCon)* (s. 1–18). IEEE. https://doi.org/10.1109/CYCON.2016.7529430

Brzozowski, R. (2021). Czynności wykonywane przez funkcjonariuszy ABW na tle zadań ABW. In P. Burczaniuk (Ed.), *Prawne aspekty funkcjonowania służb specjalnych na przykładzie Agencji Bezpieczeństwa Wewnętrznego*. Wydawnictwo ABW.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1-15.

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*(6), 1753-1819.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, *98*(1), 147-155.

Chesney, R., Citron, D., & Jurecic, Q. (2019). About that Pelosi video: What to do about 'cheapfakes' in 2020. *Lawfare*. https://www.lawfareblog.com/about-pelosi-video-what-do-about-cheapfakes-2020

Chodyka, M. (2012). *Algorytmy rozpoznawania logo oraz oznaczeń kategorii wiekowej przekazu telewizyjnego w sekwencji obrazów* (Doctoral dissertation, Politechnika Łódzka).

Choińska, Z. (2022). Wykorzystanie technologii rozpoznawania twarzy w Chińskiej Republice Ludowej. *Folia Iuridica Universitatis Wratislaviensis*, *11*(1), 75-87.

Ciekanowski, Z. (2021). Cyberprzestrzeń jako obszar współczesnych zagrożeń. In *Społeczeństwo jako uczestnik bezpieczeństwa* (s. 147-161). Akademia Sztuki Wojennej.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Czerwiński, J. (2022). Pożyteczni idioci kampanii dezinformacyjnych. Mechanizmy tworzenia się społeczności wirtualnych rozpowszechniających fałsz i manipulację w sieci. *Konteksty Społeczne*, *10*(1), 161-177.

Dakhil Zob, N., & Abdulazeez, A. M. (2024). Face recognition based on deep learning: A comprehensive review. *Indonesian Journal of Computer Science*, *7*(1), 15-30.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, s. 886-893). https://doi.org/10.1109/CVPR.2005.177

Darmohrai, T. (2022). Nowoczesna wojna informacyjna: metody i środki przeciwdziałania. *Media i Społeczeństwo, 17*(2), 39-43. https://doi.org/10.5604/01.3001.0016.3044

Dauber, C. E., & Winkler, C. K. (2014). *Visual propaganda and extremism in the online environment*. U.S. Army War College Press.

Dehmer, M., Borgert, S., & Emmert-Streib, F. (2021). Big data analysis in modern information warfare. *European Journal of Security Studies*.

Delcker, J. (2019). Welcome to the age of uncertainty. Deepfake videos may have destabilized national politics for the first time. *Politico Europe*.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-028

Dobek-Ostrowska, B. (2004). *Podstawy komunikowania społecznego*. Wydawnictwo Uniwersytetu Wrocławskiego.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton Ferrer, C. (2020). The DeepFake Detection Challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*. https://arxiv.org/abs/2006.07397

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (s. 2625-2634). https://doi.org/10.1109/CVPR.2015.7298878

Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (s. 5884-5888). https://doi.org/10.1109/ICASSP.2018.8462506

Doniec-Niezgoda, J. (2024). Zaprzepaszczone szanse Aktu o Usługach Cyfrowych? – Rozważania dotyczące definicji platformy internetowej – część I. *Prawo Mediów Elektronicznych. Kwartalnik Naukowy*.

Drosio, S., & Stanek, S. (2017). Big Data jako źródło informacji rozszerzające funkcjonowanie systemów wspomagania decyzji w zarządzaniu kryzysowym. *Zeszyty Naukowe Politechniki Częstochowskiej*, *26*, 107-120. https://doi.org/10.17512/znpcz.2017.2.10

Eireiner, A. V. (2020). Imminent dystopia? Media coverage of algorithmic surveillance at Berlin-Südkreuz. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.1.1459

Fayyad-Kazan, H., Hejase, A., Moukadem, I., & Kassem-Moussa, S. (2021). Verifying the audio evidence to assist forensic investigation. *Computer and Information Science*, *14*(3), 44-53. https://doi.org/10.5539/cis.v14n3p25

Fridrich, J., Soukal, D., & Lukáš, J. (2003). Detection of copy–move forgery in digital images. In *Proceedings of the Digital Forensic Research Workshop*.

Funk, C. (2021). The ethics of cheapfakes: The case of Nancy Pelosi. *Center for Media Engagement, University of Texas at Austin*.

Fussey, P., & Murray, D. (2023). *Independent report on the London Metropolitan Police Service's trial of live facial recognition technology*. Human Rights, Big Data & Technology Project, University of Essex.

Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, *38*(1), 84-117. https://doi.org/10.22329/il.v38i1.5068

Goban-Klas, T. (2009). *Media i komunikowanie masowe*. Wydawnictwo Naukowe PWN.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (Vol. 27).

Goryca, T. (2017). Polityka i strategia bezpieczeństwa narodowego Rzeczypospolitej Polskiej – wstęp do dyskusji. *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate et Educatione Civili*, *7*, 155-168.

Gowen, E. (2020). How misinformation on WhatsApp led to a deathly mob lynching in India. *The Washington Post*.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (s. 6645-6649). https://doi.org/10.1109/ICASSP.2013.6638947

Gu, S., & Rigazio, L. (2015). Towards deep neural network architectures robust to adversarial examples. In *ICLR Workshop*.

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (s. 1-6). https://doi.org/10.1109/AVSS.2018.8639163

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1-42. https://doi.org/10.1145/3236009

Gwagwa, A., & Mollema, W. J. T. (2024). How could the United Nations Global Digital Compact prevent cultural imposition and hermeneutical injustice? *Patterns*, *5*(1), 100822. https://doi.org/10.1016/j.patter.2023.100822

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (s. 2961-2969). https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (s. 770-778). https://doi.org/10.1109/CVPR.2016.90

Hill, K. (2022). The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics* (1st ed.). Auerbach Publications. https://doi.org/10.1201/9781003278290-26

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. https://arxiv.org/abs/1704.04861

Huang, H., & Li, J. (2021). Research on network communication model and network security technology through big data. In *IEEE International Conference on Computer and Communications* (s. 187-192). https://doi.org/10.1109/ICDSCA53499.2021.9650308

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, *25*(3), 493-516. https://doi.org/10.1177/1940161219900126

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint* arXiv:1602.07360.

Jain, A. (2019). Reverse image search: How and why it works. *Journal of Visual Communication and Image Representation*, *62*, 102-108.

Jędrasiak, K. (2025). Analiza cyberzagrożeń współczesnych kanałów komunikacyjnych. W: Z. Dacko-Pikiewicz, K. Szczepańska-Woszczyna, & M. Lis (Eds.), *Horyzonty sztucznej inteligencji a Przemysł 5.0* (s. 37-59). Akademia WSB.

Jin, N. (2023). Big data analysis of national security based on network traffic metadata. *IEEE*. https://doi.org/10.1109/ICDSI60108.2023.00020

Kacała, T. (2015). Dezinformacja i propaganda w kontekście zagrożeń dla bezpieczeństwa państwa. *Przegląd Prawa Konstytucyjnego*, *2*(26), 63-80. https://doi.org/10.15804/ppk.2015.02.03

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35-45. https://doi.org/10.1115/1.3662552

Kalpokas, I., & Kalpokienė, J. (2022). *Deepfakes: A realistic assessment of potentials, risks, and policy regulation*. Springer. https://doi.org/10.1007/978-3-030-93802-4

Kaur, A., Dhiman, A., & Singh, M. (2023). Comprehensive review: Security challenges and countermeasures for big data security in cloud computing. *IEEE*. https://doi.org/10.1109/IEMENTech60402.2023.10423449

Kaur, A., Hoshyar, A. N., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-024-10810-6

Kaźmierczak, D. (2017). Walka informacyjna we współczesnych konfliktach i jej społeczne konsekwencje. *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate et Educatione Civili*, *7*, 92-106.

Khalid, H., Tariq, S., Hassan, M. S. U., & Rehmani, M. H. (2021). Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. *arXiv preprint* arXiv:2108.07783. https://doi.org/10.1145/3476099.3484315

Kitler, W. (2019). System bezpieczeństwa narodowego RP – aspekty prawno-organizacyjne. *Wiedza Obronna*, *2*(290), 5-25.

Kowalski, J. (2020). Wykorzystanie analizy mowy w badaniach nad bezpieczeństwem informacyjnym. *Studia nad Komunikacją Społeczną*, *13*, 119-128.

Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., & Choraś, M. (2023). When explainability turns into a threat: Using xAI to fool a fake news detection method. *Computers & Security*, *126*, 103026. https://doi.org/10.1016/j.cose.2023.103599

Kozubowski, J. (2021). Socjologia wiedzy społeczeństwa cyfrowego – „praca cyfrowa" a ideologia mediów społecznościowych. *Nierówności Społeczne a Wzrost Gospodarczy*, *65*(1), 175-186. https://doi.org/10.15584/nsawg.2021.2.5

Kress, G., & van Leeuwen, T. (2020). *Reading images: The grammar of visual design* (3rd ed.). Taylor & Francis. https://doi.org/10.4324/9781003099857

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications. https://doi.org/10.4135/9781071878781

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (s. 1097-1105).

Kubit, M. (2021). Rozwój sztucznej inteligencji w świetle prawa Unii Europejskiej – dylematy i wyzwania. *Think & Make*, *3*, 33-50.

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*.

Kurek, J. (2021). Bezpieczeństwo państwa w warunkach hybrydowej regulacji danych osobowych w dobie analizy Big Data. *Akademia Sztuki Wojennej (ASzWoj)*.

Laje, D. (2023). China's deep fake law is fake. *Signal Magazine (AFCEA International)*. https://www.afcea.org/signal-media

Lasota-Kapczuk, A. (2021). *Sztuczna inteligencja w kształtowaniu bezpieczeństwa państwa w II dekadzie XXI wieku* (Doctoral dissertation, UPH w Siedlcach).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI-created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)* (s. 1-7). https://doi.org/10.1109/WIFS.2018.8630787

Li, Y., & Kuo, C.-C. J. (2003). *Video content analysis using multimodal information*. Springer. https://doi.org/10.1007/978-1-4757-3712-7

Liebowitz, D. (2018). Social media and social cleansing, Facebook and the Rohingya crisis. *Signal Magazine*. https://www.afcea.org

Liedel, K. (2011). Bezpieczeństwo informacyjne państwa. W: K. Liedel (red.), *Transsektorowe obszary bezpieczeństwa narodowego* (s. 45-62). Difin.

Liedel, K. (2010). *Zarządzanie informacją w walce z terroryzmem*. Wydawnictwo Trio.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., et al. (2015). Microsoft COCO: Common objects in context. *arXiv preprint* arXiv:1405.0312. https://doi.org/10.1007/978-3-319-10602-1_48

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation: Evidence from two randomized controlled trials. *Proceedings of the National Academy of Sciences*, *114*(10), 3887–3892.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Loy, C. C., Lin, D., Ouyang, W., Xiong, Y., et al. (2019). WIDER face and pedestrian challenge 2018: Methods and results. *arXiv preprint* arXiv:1901.08702.

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*. https://doi.org/10.1177/1365712718807226

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019). Do GANs leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval*. https://doi.org/10.1109/MIPR.2019.00103

Marvasti-Zadeh, S. M., Cheng, L., Ghanei-Yakhdan, H., & Kasaei, S. (2022). Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2020.3046478

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes. In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*. https://doi.org/10.1109/WACVW.2019.00020

Mazurek, G. (2023). Sztuczna inteligencja, prawo i etyka. *Krytyka Prawa, 15(1), 7-10*. https://doi.org/10.7206/kp.2080-1084.567

Michaelson, R. (2023). Turkish presidential candidate quits race after release of alleged sex tape. *The Guardian*. https://www.theguardian.com

Miłkowski, T. M. (2014). Czy ABW powinna korzystać ze „specjalnego" prawa do informacji? *Przegląd Bezpieczeństwa Wewnętrznego*, *6*(1), 211-226.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54(1), 1-41*. https://doi.org/10.1145/3425780

Mróz-Jagiełło, A. (2023). Rola urzędu prezydenta Rzeczypospolitej Polskiej w systemie bezpieczeństwa narodowego. *Problemy Techniki Uzbrojenia. 167(5), 79-95.* https://doi.org/10.5604/01.3001.0054.2996

Nadal, L. de, & Jančárik, P. (2024). Beyond the deepfake hype: AI, democracy, and "the Slovak case". *HKS Misinformation Review*.

Naik, N., Kumar, P., Aggarwal, R., & Upadhyay, S. (2023). AI-generated voice deepfakes and security implications: A systematic review. *Journal of Cybersecurity and Privacy*.

Neuendorf, L. A. (2016). *The content analysis guidebook* (2nd ed.). SAGE Publications. https://doi.org/10.4135/9781071802878

Nguyen, L., & Chen, M. (2021). Early detection of cyber threats: Big data analytics approaches. *Cyber Security International Journal*.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection: A survey. *arXiv preprint* arXiv:1909.11573.

Niedzwiedskaja, N. P. (2019). Społeczeństwo cyfrowe: problemy i perspektywy rozwoju. W: J. M. Osipow & A. Z. Nowak (red.), *Rewolucja cyfrowa. Wyzwania, problemy, perspektywy rozwoju*. Wydawnictwo WZ UW.

Nowak, E., & Nowak, M. (2011). *Zarys teorii bezpieczeństwa narodowego*. Difin.

Nowakowski, Z., Pomykała, M., & Rajchel, J. (2009). Pojęcie bezpieczeństwa i porządku publicznego. W: K. Rajchel (Ed.), *Administracja bezpieczeństwa i porządku publicznego*. Towarzystwo Naukowe Powszechne.

Nowicka, J., Pauliuchuk, Y., Ciekanowski, Z., Falda, B., & Sikora, K. (2024). The use of data analytics in human resource management. *European Research Studies Journal, XXVII(2), 203-215*. https://doi.org/10.35808/ersj/3380

Nowicka, J., & Załoga, W., & Ciekanowski, Z. (2018). Komunikacja strategiczna w naukach o zarządzaniu i jakości oraz w naukach o bezpieczeństwie. *Zeszyty Naukowe WSZOP w Katowicach*.

Nowiński, M., & Wieczerza, M. (2021). Zadania służb specjalnych w obszarze ochrony informacji niejawnych ze szczególnym uwzględnieniem aspektów międzynarodowych. In P. Burczaniuk (Ed.), *Prawne aspekty funkcjonowania służb specjalnych na przykładzie ABW*. Wydawnictwo ABW.

Nye, J. (2004). *Soft power: The means to success in world politics*. PublicAffairs.

Ogonowska, A. (2021). Od intrygi i spisku do zagrożeń hybrydowych. Nowe kompetencje i edukacja medialna w cyberkulturze 3.0. *Studia de Cultura, 14(1), 7-22*. https://doi.org/10.24917/20837275.14.1.1

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation. *Human Factors*, *52*(3), 381-410. https://doi.org/10.1177/0018720810376055

Patino, L., Caicedo, J., Arguello, R., & Gomez, H. (2022). Speaker recognition approach using mel-frequency cepstral coefficients and deep neural networks. *IEEE Access*.

Pellerin, C. (2017). Project Maven to deploy computer algorithms to war zone by year's end. *U.S. Department of Defense*.

Peterson, S., Jackson, B. A., Woods, D., & Treyger, E. (2023). Finding a broadly practical approach for regulating the use of facial recognition by law enforcement. *RAND Corporation*.

Piccardi, M. (2004). Background subtraction techniques: A review. In *IEEE International Conference on Systems, Man and Cybernetics* (s. 3099-3104). https://doi.org/10.1109/ICSMC.2004.1400815

Piękoś, K. (2023). Organizacja systemu zarządzania kryzysowego w Polsce. *Annales Scientia Politica*.

Pinhasov, B., Lapid, R., Ohayon, R., Sipper, M., & Aperstein, Y. (2024). Attacks on deepfake detectors. Afeka – The Academic College of Engineering, Tel Aviv / Ben-Gurion University of the Negev / DeepKeep.

Pokruszyński, W. (2009). *Współczesne bezpieczeństwo narodowe*. WSGE.

Prońko, J., & Kaczmarczyk, B. (2015). Bezpieczeństwo państwa – pojęcie, istota, strategia, system. *Rocznik Bezpieczeństwa Morskiego*.

Ptaszek, G. (2019). *Edukacja medialna 3.0. Krytyczne rozumienie mediów cyfrowych w dobie Big Data i algorytmizacji*. Wydawnictwo Uniwersytetu Jagiellońskiego.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. https://doi.org/10.1109/5.18626

Rajeshwari, R. (2019). Mob lynching and social media. *Yale Journal of International Affairs*.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint*. https://arxiv.org/abs/1804.02767

Reimao, R., & Tzerpos, V. (2021). Synthetic speech detection using neural networks. In *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. https://doi.org/10.1109/SpeD53181.2021.9587406

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint* https://arxiv.org/abs/1506.01497. https://doi.org/10.48550/arXiv.1506.01497

Roozenbeek, J., & Van der Linden, S. (2019). Fake News Game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-8. https://doi.org/10.1057/s41599-019-0279-9

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (s. 1-11). https://doi.org/10.1109/ICCV.2019.00009

Saldaña, M. B. (2016). *The coding manual for qualitative researchers* (3rd ed.). SAGE Publications.

Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. *J. Imaging*, *9*(3), 95. https://doi.org/10.3390/jimaging9060122

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *ITU Journal: ICT Discoveries*, *1*(1), 18-29.

Siciliano, F., Maiano, L., Papa, L., Baccini, F., Amerini, I., Silvestri, F. (2023). Adversarial data poisoning for fake news detection: How to make a model misclassify a target news without modifying it. *arXiv preprint*. https://arxiv.org/abs/2304.04952

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In: *Advances in Neural Information Processing Systems* (Vol. 27, s. 568-576).

Skomra, W. (2017). Miejsce, rola i zadania Rządowego Centrum Bezpieczeństwa w zakresie integracji działań na rzecz utrzymania bezpieczeństwa. W: J. Telak (red.), *Ratownictwo wodne oraz inne aspekty bezpieczeństwa na obszarach wodnych. Dylematy i wyzwania* (s. 289-307). Centrum Szkolenia Policji.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349-1380. https://doi.org/10.1109/34.895972

Stupp, C. (2019). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *Wall Street Journal*. https://www.wsj.com/articles

Sultan, D. A., & Ibrahim, L. M. (2023). Video-audio multimodal for deepfake detection. *Proceedings of the National Academy of Sciences (PNAS)*, *120*(7), e2221340120.

Swoboda, P. (2014). Agencja Bezpieczeństwa Wewnętrznego w systemie bezpieczeństwa państwa. W: M. Śliwa, A. Żebrowski, & R. Kłaczyński (red.), *Bezpieczeństwo RP. Wczoraj i dziś. Studia z zakresu bezpieczeństwa państwa* (s. 209-222). Wydawnictwo Naukowe Uniwersytetu Pedagogicznego.

Śmiałek, W. (2023). *Siły zbrojne w polityce i systemie bezpieczeństwa państwa*. Wydawnictwo Naukowe FNCE.

Tabakow, M., Korczak, J., & Franczyk, B. (2014). Big data – definicje, wyzwania i technologie informatyczne. *Informatyka Ekonomiczna*, *18*(3), 6-17.

Taczkowska-Olszewska, J. (2017). Bezpieczeństwo informacyjne jako kategoria prawna. Ujęcie teoretyczne. W: W. Kitler & J. Taczkowska-Olszewska (red.), *Bezpieczeństwo informacyjne. Aspekty prawno-administracyjne* (s. 45-58). Towarzystwo Wiedzy Obronnej.

Takahashi, N., Gygli, M., Pfister, B., & Van Gool, L. (2016). Deep convolutional neural networks and data augmentation for acoustic event detection. W: *Interspeech* (s. 1-5). https://doi.org/10.21437/Interspeech.2016-805

Tandoc, E. C., Lim, Z. W., & Ling, R. (2017). Defining fake news: A typology of scholarly definitions. *Digital Journalism*, *5*(2), 210-232. https://doi.org/10.1080/21670811.2017.1360143

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (s. 4489-4497). https://doi.org/10.1109/ICCV.2015.510

Tsai, Y. H. H., Bai, S., & Morency, L. P. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (s. 1482-1491). https://doi.org/10.18653/v1/P19-1656

Turchini, F., Seidenari, L., Uricchio, T., & Del Bimbo, A. (2018). Deep learning-based surveillance system for open critical areas. *Inventions*, *3*(4), 54. https://doi.org/10.3390/inventions3040069

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE*, *18*(6), e0270651. https://doi.org/10.21428/cb6ab371.2a97c514

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(2), 1-12. https://doi.org/10.1177/2056305120903408

Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: A review. *IEE Proceedings – Vision, Image & Signal Processing*, *152*(2), 1-13. https://doi.org/10.1049/ip-vis:20041147

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (s. 511-518). https://doi.org/10.1109/CVPR.2001.990517

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wachowicz, M. J. (2019). Ujęcie teoretyczne pojęcia dezinformacji. *Wiedza Obronna*, *1*(3), 67-85.

Waltzman, R. (2017). *The weaponization of information: The need for cognitive security*. RAND Corporation. https://doi.org/10.7249/CT473

Wang, Y., & Chung, S. H. (2022). Artificial intelligence in safety-critical systems: A systematic review. *Industrial Management & Data Systems*, *122*(10), 1877-1900. https://doi.org/10.1108/IMDS-07-2021-0419

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

Wasiuta, O., & Wasiuta, S. (2019). Deepfake jako skomplikowana i głęboko fałszywa rzeczywistość. *Annales Universitatis Paedagogicae Cracoviensis. Folia 289, Studia de Securitate*, *14*(2), 45-58.

Wąsik, K. (2022). Bezpieczeństwo informacyjne państwa. In *13th Eurasian Conference on Language and Social Sciences, Daugavpils University (Łotwa)*, 220-230.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 13-22. https://doi.org/10.22215/timreview/1282

Wieczorkowski, J. (2017). Big data a prywatność. Naruszenie prywatności w świecie wirtualnym. *Roczniki KAE*, *10*(3), 1-12.

Wieczorkowski, J. (2018). Metody big data w zapewnieniu bezpieczeństwa publicznego a problem prywatności. *Roczniki KAE*, *11*(1), 45-58.

Więcaszek-Kuczyńska, L. (2014). Zagrożenia bezpieczeństwa informacyjnego. *Obronność – Zeszyty Naukowe WZiD ASzWoj*, *6*, 1-10.

Wigienka, S. (2021). Przeciwdziałanie dezinformacji w dobie pandemii COVID-19. *Com.press*, *18*(4), 39-46. https://doi.org/10.51480/compress.2021.4-1.247

Wiśniewski, B. B. (2013). *System bezpieczeństwa państwa. Konteksty teoretyczne i praktyczne*. WSPol.

Witczak, M. (2024). Biały wywiad w zarządzaniu bezpieczeństwem informacji. *Przegląd Bezpieczeństwa Wewnętrznego*, *23*(1), 33-47. https://doi.org/10.4467/20801335PBW.24.009.19611

Wróblewski, T. (2024). Sztuczna inteligencja jako narzędzie do walki z dezinformacją. *IJOLS (Poznań)*, *17*(1), 15-29. https://doi.org/10.5604/01.3001.0054.6967

Yadav, P., Sharma, A., Jadon, S., & Murala, S. (2021). Deepfake detection using InceptionResNetV2 and LSTM. In *International Conference on Emerging Technologies* (s. 134-139).

Yang, M., Liao, W., Yang, C., Cao, Y., & Rosenhahn, B. (2018). Security event recognition for visual surveillance. *arXiv preprint*. https://arxiv.org/abs/1806.10512

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys*, *38*(4), 1-45. https://doi.org/10.1145/1177352.1177355

Zakrzewski, P. (2023). Analiza sieci społecznych w badaniach nad bezpieczeństwem militarnym i niemilitarnym państwa. *Zeszyty Naukowe WSOWL*, *16*(1), 22-40.

Zdrodowski, B. (2019). Istota bezpieczeństwa państwa. *Annales Universitatis Paedagogicae Cracoviensis. Studia de Securitate*, *13*, 77-89. http://hdl.handle.net/11716/11892

Zhang, T., Liu, Z., Lian, X., & Wang, X. (2010). Study on moving-objects detection technique in video surveillance system. In *Chinese Control and Decision Conference* (s. 2065-2069). https://doi.org/10.1109/CCDC.2010.5498533

Zhang, Y., Ma, X., Wang, X., & Sun, J. (2021). Detecting audio deepfakes using frame-level self-attention. *IEEE Transactions on Dependable and Secure Computing*, *18*(1), 123-135. https://doi.org/10.1109/TDSC.2020.2982720

Zhou, Y., & Lim, S.-N. (2021). Joint audio-visual deepfake detection. In *IEEE International Conference on Computer Vision (ICCV)* (s. 4730-4740). https://doi.org/10.1109/ICCV48922.2021.00464

Wykaz rysunków

Rysunek 1. Uniwersalny model systemu bezpieczeństwa narodowego..... 20
Rysunek 2. Schemat przetwarzania treści audiowizualnych w czasie rzeczywistym..... 109
Rysunek 3. Schemat architektury systemu analizy wideo w czasie rzeczywistym na potrzeby bezpieczeństwa operacyjnego 130
Rysunek 4. Przykład działania inteligentnego systemu monitoringu miejskiego – automatyczne wykrywanie i śledzenie osób w przestrzeni publicznej z wykorzystaniem algorytmów sztucznej inteligencji 159
Rysunek 5. Porównanie skuteczności badanych metod analizy audiowizualnej w detekcji manipulacji – zależność między czułością a specyficznością 179

Wykaz tabel

Tabela 1. Typowe techniki tworzenia fake news i manipulacji audiowizualnej 75
Tabela 2. Przykłady regulacji krajowych dotyczących monitorowania treści audiowizualnych 127
Tabela 3. Stanowiska wybranych organizacji międzynarodowych w sprawie dezinformacji i *deepfake* 127
Tabela 4. Klasyczne a nowoczesne metody analizy wideo – kluczowe różnice 136
Tabela 5. Charakterystyka poziomów integracji 144
Tabela 6. Wyniki eksperymentalnych testów skuteczności algorytmów detekcji manipulacji audiowizualnych 178
Tabela 7. Wyniki eksperymentalne – efekty prewencyjnej komunikacji opartej na teorii inoculacji 202
Tabela 8. Porównanie kompetencji kluczowych w 2015 i 2025 roku 222

ISBN 978-83-68103-18-2