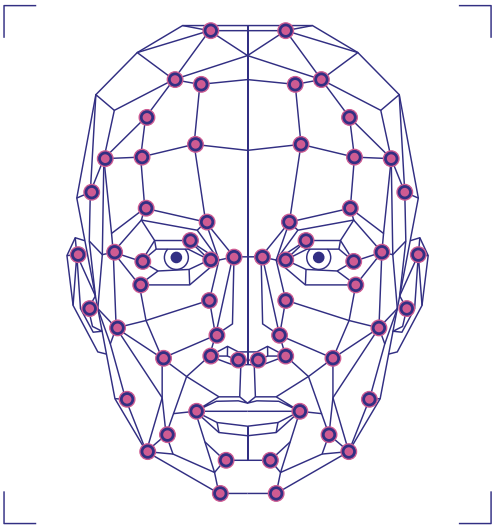
POLICY PAPER

Deep fakes i nadużycia na tle seksualnym – prawo wobec wyzwań syntetycznych mediów

Obserwatorium
Demokracji Cyfrowej



POLICY PAPER

Deep fakes i nadużycia na tle seksualnym – prawo wobec wyzwań syntetycznych mediów

Obserwatorium
Demokracji Cyfrowej

Fundacja Obserwatorium Demokracji Cyfrowej
Kontakt: info@ceeddw.org

Autorzy:

Mateusz Łabuz jest badaczem w Institut für Friedensforschung und Sicherheitspolitik an der Universität Hamburg (IFSH) i doktorantem na Technische Universität Chemnitz. Przez siedem lat był zawodowym dyplomatą w Ministerstwie Spraw Zagranicznych RP. Wykłada cyberbezpieczeństwo, sztuczną inteligencję, dezinformację i fact-checking na Uniwersytecie Komisji Edukacji Narodowej i Uniwersytecie Papieskim Jana Pawła II w Krakowie. Jego główne zainteresowania badawcze to media syntetyczne, ze szczególnym uwzględnieniem deep fakes, budowania odporności społecznej, a także zagrożeń kognitywnych i hybrydowych. Opublikował liczne analizy poświęcone regulowaniu i definiowaniu deep fakes w Akcie o Sztucznej Inteligencji.
ORCID: 0000-0002-6065-2188

Dr hab. Mikołaj Małecki jest doktorem habilitowanym nauk prawnych, adiunktem w Katedrze Prawa Karnego Uniwersytetu Jagiellońskiego, prezesem Krakowskiego Instytutu Prawa Karnego Fundacja. Autor kilkuset opracowań poświęconych prawu karnemu. Laureat stypendium Ministra Nauki i Szkolnictwa Wyższego dla wybitnych młodych naukowców (2019-2022). Dwukrotnie wyróżniany w rankingach Dziennika Gazety Prawnej na najbardziej wpływowych prawników w Polsce. Popularyzator nauki, autor bloga i portalu DogmatyKarnisty.pl.
ORCID: 0000-0002-2878-2791

Adw. dr Karolina Mania jest adiunktem w Instytucie Ekonomii, Finansów i Zarządzania na Wydziale Zarządzania i Komunikacji Społecznej Uniwersytetu Jagiellońskiego specjalizującym się w tematyce legal tech, online dispute resolution, deep fake oraz handlu elektronicznego. Autorka publikacji z zakresu sporów dotyczących domen internetowych, elektronicznych metod rozwiązywania sporów, narzędzi IT dla prawników oraz specyfiki rynku usług prawnych. Od 2019 roku pełni funkcję prezesa stowarzyszenia Klub Stypendystów Fundacji Kościuszkowskiej zrzeszającego beneficjentów programów stypendialnych nowojorskiej Fundacji Kościuszkowskiej.
ORCID: 0000-0001-9063-7563

Deep fakes, syntetyczne media w formie dźwiękowej lub wizualnej, generowane przy użyciu technologii sztucznej inteligencji, stanowią istotne wyzwanie dla współczesnego prawa. Ich potencjał do tworzenia realistycznych, lecz nieprawdziwych treści wykracza daleko poza kwestie dezinformacji politycznej, stając się narzędziem szeroko stosowanym w celach oszustw finansowych czy nadużyć seksualnych. W szczególności szkodliwe i problematyczne, także ze względu na kwalifikację prawną, są przypadki wykorzystywania deep fakes do produkcji niekonsensualnych treści intymnych przedstawiających osoby dorosłe oraz syntetycznych materiałów prezentujących wykorzystywanie seksualne małoletnich. Celem niniejszego opracowania jest analiza zagrożeń oraz wskazanie niezbędnych rozwiązań, szczególnie w obszarze prawa karnego, umożliwiających dostosowanie istniejących przepisów do nowej rzeczywistości technologicznej.

- Rozwój generatywnej sztucznej inteligencji przyczynia się do wykształcenia nowych problemów natury prawnej, co wymaga reinterpretacji istniejących bądź tworzenia nowych adekwatnych przepisów, zwłaszcza wobec konieczności wdrażania przez Polskę regulacji o charakterze ponadnarodowym, czego przykładem jest m.in. unijna Dyrektywa w sprawie zwalczania przemocy wobec kobiet i przemocy domowej z 2024 r.

- Powszechnie dyskutowanym przejawem szkodliwości deep fakes jest ich wpływ na dezinformację o charakterze politycznym. Problematyka niekonsensualnych treści o charakterze seksualnym w formie deep fakes, w tym konsekwencje w postaci wiktymizacji kobiet i dzieci, wymaga większego zainteresowania ustawodawcy, w tym Ministerstw Cyfryzacji, Sprawiedliwości, Zdrowia, Rodziny, Pracy i Polityki Społecznej, zwłaszcza w obliczu konieczności wdrożenia wskazanej Dyrektywy do 14 czerwca 2027 r.

- Deep fakes są coraz częściej wykorzystywane do tworzenia treści pedofilskich, co prowadzi do wiktymizacji nowych ofiar, multiplikacji szkodliwych materiałów cyrkulujących online oraz istotnego obciążenia organów ścigania i biegłych mających trudności w odróżnianiu materiałów prawdziwych od syntetycznych.

- Niezbędne są konkretne zmiany w prawie, które umożliwią lepszą ochronę najbardziej wrażliwych grup i przeciwdziałanie negatywnym skutkom społecznym – powinny one obejmować konkretne przepisy penalizujące tworzenie i udostępnianie niekonsensualnych treści intymnych prezentujących osoby dorosłe (np. poprzez modyfikację art. 191a k.k.) oraz dostosowanie istniejących już przepisów Kodeksu karnego w zakresie, w jakim penalizują one produkowanie, rozpowszechnianie, prezentowanie, przechowywanie lub posiadanie treści pornograficznych przedstawiających wytworzony albo przetworzony wizerunek małoletniego uczestniczącego w czynności seksualnej (art. 204 par. 4b k.k.).

- Odpowiedź na poziomie państwowym powinna uwzględniać zarówno bezpośrednie zmiany w prawie karnym, jak i inwestowanie w dodatkowe mechanizmy ochrony obejmujące zapewnienie ofiarom odpowiedniej pomocy psychologicznej i prawnej, edukację i programy podnoszące świadomość wśród dzieci i dorosłych, jak również zwiększenie nakładów na szkolenia pracowników wymiaru sprawiedliwości i wykorzystywane przez nich technologie.

WPROWADZENIE

Dynamiczny rozwój sztucznej inteligencji (AI) jest katalizatorem postępu oraz rozwoju w licznych obszarach życia społecznego i gospodarki. Koszty społeczne rewolucji cyfrowej są aktualnie trudne do kwantyfikowania ze względu na złożoność i innowacyjność tych procesów. Holistyczne podejście do rozwoju technologicznego musi także uwzględniać dostrzegalne negatywne trendy, co z jednej strony umożliwi dynamiczne reagowanie (także w obszarze prawnym), a z drugiej będzie elementem budowania społecznego zaufania do technologii, której ewolucja powinna respektować podstawowe zasady etyczne.

W ostatniej dekadzie liczne organizacje wypracowały podstawy etyki dla rozwijania technologii AI. Znajduje to odzwierciedlenie w uznawanych za jeden ze standardów zasadach OECD obejmujących m.in. tworzenie AI godnej zaufania, odpowiedzialnej, inkluzywnej, niedyskryminującej, projektowanej z poszanowaniem praworządności, praw człowieka oraz wartości demokratycznych[1].

Rozwój generatywnych modeli AI (w tym generatywnych sieci kontradyktoryjnych) i upowszechnienie dostępu do technologii sprawiły, że tworzenie syntetycznych mediów (generowanych z udziałem AI) stało się łatwiejsze i bardziej powszechne. To zjawisko jest często określane mianem „demokratyzacji"[2]. Obsługa wyrafinowanych narzędzi, do niedawna zarezerwowanych wyłącznie dla ekspertów, aktualnie nie wymaga zaawansowanych umiejętności i wiedzy technicznej, co zwiększa ich dostępność, ale i skalę potencjalnych zagrożeń.

Deep fakes to jedna z form syntetycznych mediów, generowanych przy użyciu AI z wykorzystaniem metod uczenia maszynowego. Technologia ta pozwala stworzyć realistyczne wideo, obrazy lub audio, które mogą przedstawiać osoby w sytuacjach, które nigdy nie miały miejsca. Choć w wielu wypadkach są używane do celów artystycznych i badawczych, ich zastosowania obejmują także obszar dezinformacji, cyberprzestępczości i nadużyć na tle seksualnym[3].

Deep fakes nie mają utrwalonej i uniwersalnej definicji prawnej. Polskie ustawodawstwo powinno się posługiwać definicją wypracowaną w unijnym Akcie o Sztucznej Inteligencji (AI Act) przyjętym w 2024 r. W Artykule 3(60) AI Act deep fakes zostały opisane jako: „wygenerowane przez AI lub zmanipulowane przez AI obrazy, treści dźwiękowe lub treści wideo, które przypominają istniejące osoby, przedmioty, miejsca, podmioty lub zdarzenia, które odbiorca mógłby niesłusznie uznać za autentyczne lub prawdziwe"[4]. Definicja uwypukla zatem manipulacyjny charakter syntetycznych treści i ich potencjał do wprowadzania odbiorców w błąd[5].

1 OECD AI Policy Observatory (2024). *OECD AI Principles overview*. https://oecd.ai/en/ai-principles.

2 Zob. m.in. Farid H., Schindler H.-J. (2020). *Deep fakes. On the Threat of Deep fakes to Democracy and Society*. Konrad Adenauer Stiftung: Berlin.

3 Zob. m.in. Vaccari C., Chadwick A. (2021). *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*. „Social Media + Society". Nr 6(1); van Huijstee M. i in. (2021). *Tackling deepfakes in European policy*. European Parliamentary Research Service: Bruksela.

4 Artificial Intelligence Act (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence*.

5 Zob. Łabuz M. (2025). *A Teleological Interpretation of the Definition of DeepFakes in the EU Artificial Intelligence Act—A Purpose-Based Approach to Potential Problems With the Word „Existing"*. „Policy & Internet".

Już pierwsze przypadki tworzenia deep fakes w 2017 r. powinny były stać się wyraźnym sygnałem ostrzegawczym. Technologia była bowiem wykorzystywana do tworzenia niekonsensualnych treści pornograficznych, w których wizerunek osób publicznych (najczęściej celebrytek) był nakładany na wcześniej zarejestrowane materiały pornograficzne, tworząc mylne wrażenie uczestnictwa w tego typu nagraniach[6]. Od 2017 r. technologia zanotowała istotny postęp jakościowy. Współcześnie, czego dowodzą liczne badania, syntetyczne media są dla przeciętnego odbiorcy nieodróżnialne od mediów prawdziwych[7]. Deep fakes zacierają tym samym granicę między rzeczywistością a fikcją, podważając fundamenty zaufania społecznego do mediów i informacji, a tym samym paradygmat wiary w to, co widzimy. Jednocześnie pozwalają na wierne imitowanie rzeczywistości, czego konsekwencje są odczuwalne w świecie prawdziwym.

Deep fakes stwarzają liczne wyzwania w wielu obszarach. Są m.in. wykorzystywane do wspierania dezinformacji, co stereotypowo łączone jest głównie z potencjałem do manipulacji wyborczych. Ich szkodliwe zastosowania powinny być jednak rozpatrywane znacznie szerzej. Nie tylko umożliwiają one tworzenie treści fałszywych, które mogą zniszczyć reputację jednostek, ale również są wykorzystywane w cyberprzemocy, w tym szantażach, wymuszeniach oraz szerzeniu przemocy na tle seksualnym, prowadząc do daleko idących konsekwencji społecznych i politycznych, a przede wszystkim wywołując negatywne skutki psychologiczne dla poszkodowanych, z których istotną grupę stanowią kobiety. Stwarzają także nowe wyzwania w zakresie prawa karnego, cywilnego czy autorskiego, jak również ochrony prywatności oraz praw człowieka. Tymczasem brak jednoznacznych regulacji prawnych wobec tworzenia i rozpowszechniania deep fakes o charakterze seksualnym, nieprecyzyjność istniejących regulacji lub niekonkretność ich interpretacji w obliczu rozwoju nowoczesnych technologii utrudniają ściganie działań odznaczających się wysokim stopniem społecznej szkodliwości.

Obecne regulacje w większości krajów UE nie regulują wprost deep fakes, a w takich wypadkach istniejące przepisy prawa karnego lub cywilnego mogą być wykorzystywane subsydiarnie. Wobec niewielkiej liczby precedensowych spraw wykładnia przepisów dostosowana do nowej rzeczywistości technologicznej nie jest ugruntowana[8]. Próby wprowadzania adekwatnych rozwiązań w wybranych krajach, w tym w Wielkiej Brytanii w postaci Online Safety Act, mogą być istotnym punktem odniesienia, podobnie jak dyskusje towarzyszące formułowaniu przepisów[9].

Korzystanie z rozwiązań, które były tworzone dla „świata analogowego"[10], w kontekście przestępstw seksualnych, oznacza, że ofiary mierzą się z traumą wynikającą z dystrybucji materiałów przedstawiających ich wizerunek w sytuacjach intymnych, muszą udowadniać, że takie materiały są fałszywe, a w przestrzeni prawnej będą mieć trudności z wyegzekwowaniem sprawiedliwości ze względu na niejednoznaczność przepisów lub wręcz brak możliwości kwalifikacji prawnej[11]. Problem staje się jeszcze bardziej dotkliwy w przypadku dziecięcej pornografii, gdzie identyfikacja fałszywych treści wymaga zaawansowanych tech-

6 Brooks T. i in. (2019). *Increasing Threats of Deepfake Identities*. Department of Homeland Security: Waszyngton.

7 Zob. m.in. Nightingale S. J., Farid H. (2022). *AI-synthesized faces are indistinguishable from real faces and more trustworthy*, „PNAS". Nr 119(8).

8 Zob m.in. Ziobroń A. (2021). *Deepfake a prawo karne. Uwagi de lege lata i de lege ferenda dotyczące fałszywej pornografii.* „Studenckie Prace Prawnicze, Administratywistyczne i Ekonomiczne". Nr 37; Vera G. G. (2024). Deep fake – postęp technologiczny a prawo karne. „Acta Iuridica Resoviensia". Nr 1(44).

9 Zob. m.in. McGlynn C., Toparlak R. (2024). *Creating Sexually Explicit Deepfakes: Options for Criminal Law Reform.* Hate Aid & Durham University.

10 Sittig J. (2024). *Strafrecht und Regulierung von Deepfake-Pornografie.* Bundeszentrale für politische Bildung: Berlin.

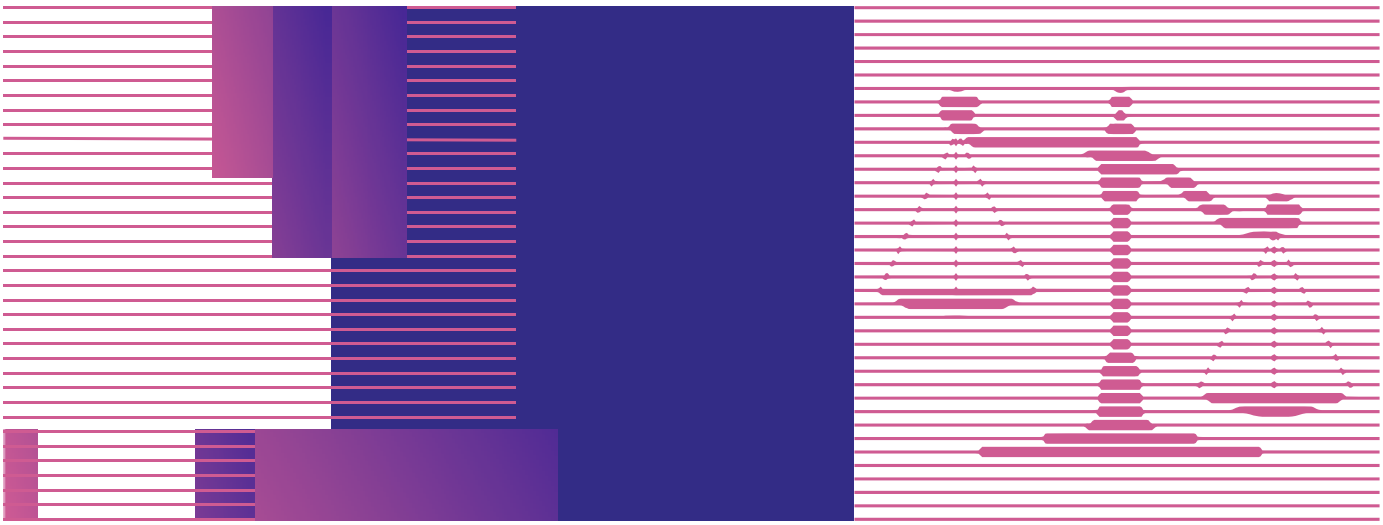
11 Okolie C. (2023). *Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns.* „Journal of International Women's Studies". Nr 25(2).

nologicznie narzędzi oraz wiedzy eksperckiej, a ich udostępnianie prowadzi do multiplikacji szkodliwych treści cyrkulujących online[12].

W miarę nasilania się problemu wzrasta potrzeba wypracowania regulacji chroniących ofiary oraz zapobiegających nadużyciom. Wyzwaniem jest zapewnienie równowagi między efektywnością regulacji a ochroną wolności słowa oraz prawa do prywatności, które jednak nie mogą być pretekstem do niepodejmowania działań ochronnych.

Niniejsze opracowanie przedstawia konkretne rozwiązania na gruncie prawa karnego, które powinny być punktem wyjścia do prac legislacyjnych. Jest to szczególnie ważne w obliczu konieczności wdrażania Dyrektywy w sprawie zwalczania przemocy wobec kobiet i przemocy domowej. Wobec istotności wyzwań opisane problemy prawne powinny zostać rozwiązane zdecydowanie wcześniej niż przewidziany w Dyrektywie termin czerwca 2027 r. Nasze rekomendacje mogą posłużyć do utrwalenia nowej linii interpretacyjnej już istniejących przepisów dostosowanej do dynamicznego rozwoju technologicznego, bądź też do stworzenia nowych bardziej adekwatnych i jednoznacznych przepisów zapewniających wyższy poziom ochrony poszkodowanych i służących wzmacnianiu społecznej świadomości. Propozycje są podyktowane chęcią zwiększania zaufania do porządku prawnego i przeciwdziałania ewentualnej arbitralności przyszłych rozstrzygnięć nieosadzonych dotychczas w konkretnej linii orzeczniczej. W wymiarze prawnym mają służyć zwiększeniu precyzyjności i konkretności interpretacji. Nadrzędnym celem jest zabezpieczenie praw ofiar i przeciwdziałanie niewłaściwemu wykorzystaniu technologii, co w wymiarze społecznym i politycznym może także służyć przeciwdziałaniu dyskryminacji na tle płciowym, zapobieganiu uprzedmiatawiania kobiet oraz wzmocnieniu odporności systemów demokratycznych.

12 Internet Watch Foundation (2023). *How AI is being abused to create child sexual abuse imagery.* Internet Watch Foundation: Cambridge.



NSII w formie deep fakes – problematyka i propozycje działań na gruncie prawa karnego

Jednym z kluczowych zagrożeń o charakterze społecznym i politycznym związanych z wykorzystywaniem deep fakes jest tworzenie niekonsensualnych treści intymnych prezentujących osoby dorosłe. Powszechnie opisuje się je jako *deep porn*, co semantycznie nawiązuje do połączenia terminów *deep fakes* i *pornography*, jednak część badaczy sugeruje używanie bardziej adekwatnego terminu *non-consensual synthetic intimate imagery* (NSII), który zwraca uwagę przede wszystkim na aspekt abuzywny – niekonsensualne wykorzystanie cudzego wizerunku i obrazów intymnych[13]. Warto w tym wypadku odnotować, iż obrazy służące za podstawę syntezy AI mogą być pozyskiwane przez sprawców z dostępnych publicznie źródeł, a brak konsensualności odnosi się przede wszystkim do braku zgody na wykorzystanie wizerunku. W dalszej części opracowania, ze względu na precyzyjność przekazu, stosowane jest sformułowanie „NSII w formie deep fakes".

NSII jest dominującym przejawem wykorzystania deep fakes w formie wideo – stanowi ponad 90% z nich[14]. W blisko 100% przypadków NSII w formie deep fakes dotyka kobiet, co prowadzi do wiktymizacji tysięcy z nich na całym świecie, a także przyczynia się do wzmacniania stereotypów płciowych oraz uprzedmiotowienia kobiet, wywierając trudny do skwantyfikowania wpływ na społeczeństwo.

NSII stało się przedmiotem zainteresowania licznych badaczy, a raporty poświęcone temu negatywnemu zjawisku potwierdzają istotny trend związany z płcią ofiar oraz łączeniem przemocy na tle seksualnym z atakami motywowanymi ideologicznie bądź politycznie[15]. Jednym z nadrzędnych celów wykorzystania NSII jest „upokarzanie, zawstydzanie i traktowanie kobiet przedmiotowo, zwłaszcza tych, które mają odwagę głośno wyrażać swoje poglądy"[16]. Opublikowany w grudniu 2024 r. raport dotyczący NSII w formie deep fakes wymierzonych w członków Kongresu USA wykazał, iż na 26 rozpoznanych przypadków, aż 25 dotyczyło kobiet[17]. Ofiarami NSII w formie deep fakes padały także prominentne polityczki i działaczki społeczne, co z jednej strony można łączyć z ich ekspozycją medialną i popularnością, a z drugiej z prezentowanymi poglądami. Już sam fakt możliwości targetowania ofiar na bazie ich aktywności może być ważnym czynnikiem zniechęcającym kobiety do zaangażowania w przestrzeni publicznej, negatywnie wpływając na kształtowanie demokratycznego społeczeństwa i równouprawnienia.

13 Zob. Viola M., Voto C. (2023). *Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images.* „Synthese". Nr 201(1); Rigotti C., McGlynn C. (2022). *Towards an EU criminal law on violence against women: The ambitions and limitations of the Commission's proposal to criminalise image-based sexual abuse.* „New Journal of European Criminal Law". Nr 13(4).

14 Home Security Heroes (2023). *2023 State of Deepfakes.* https://www.homesecurityheroes.com/state-of-deepfakes. Zob. także wcześniejsze badania: Ajder i in. (2019). *The State of Deepfakes: Landscape, Threats, and Impact.* Deeptrace: Amsterdam.

15 Maddocks S. (2020). *'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deep fakes.* „Porn Studies". Nr 7(4); Rosalie Li E., Schultz B., Jankowicz N. (2024). *Deepfake Pornography Goes o Washingon: Measuring the Prevalence of AIGeneraed NonConsensual Intimae Imagery Targeting Congress.* American Sunlight Project: Waszyngton.

16 Jankowicz N. (2023). *I Shouldn't Have to Accept Being in Deepfake Porn.* https://www.theatlantic.com/ideas/archive/2023/06/deepfake-porn-ai-misinformation/674475.

17 Rosalie Li i in., *op. cit*

Skala problemu NSII w formie deep fakes nie jest w pełni zbadana. Znaczna część zgłoszeń nadużyć ginie w policyjnych statystykach, wiele z przypadków w ogóle nie jest zgłaszanych przez ofiary, co wynika z odczuwanego przez nie wstydu, upokorzenia czy bezradności oraz niskiej świadomości społecznej. Wymienione elementy regularnie pojawiają się w zeznaniach ofiar, podobnie jak obawy przed wtórną wiktymizacją oraz ostracyzmem społecznym. Tymczasem konsekwencje ataków na integralność cielesną pozostawiają trwałe ślady o charakterze psychologicznym. Ofiary zmagają się depresją, stanami lękowymi, a w skrajnych przypadkach dochodzi do prób samobójczych[18].

„Javellana stawała się coraz bardziej lękliwa i paranoiczna. Przestała wychodzić sama nocą, zaczęła wielokrotnie sprawdzać, czy drzwi i okna są zamknięte zanim poszła spać. Aby chronić swoje życie osobiste, oznaczyła swój profil na Instagramie jako prywatny i usunęła swoje zdjęcia w kostiumie kąpielowym. Uczestnictwo w konferencjach prasowych było częścią jej pracy, ale teraz czuła niepokój za każdym razem, gdy ktoś podnosił aparat, by zrobić zdjęcie. Martwiła się, że jej publiczne zdjęcia zostaną przekształcone w pornografię" – fragment artykułu Coralie Kraft poświęconego amerykańskiej urzędniczce Sabrinie Javellana, która padła ofiarą NSII w formie deep fakes[19].

Kraje takie jak USA czy Korea Południowa mierzą się z istotną liczbą przypadków niekonsensualnej syntetycznej przemocy seksualnej, a negatywne trendy mają istotny potencjał do replikacji w innych państwach. Jest to związane ze wspomnianą wcześniej „demokratyzacją" dostępu do technologii i znaczącym uproszczeniem mechanizmów tworzenia syntetycznych mediów, którym nie towarzyszy wypracowanie odpowiednich systemów zabezpieczających przed nadużyciami.

Dodatkowym problemem rozpoznanym w licznych krajach jest posługiwanie się przez dzieci i młodzież aplikacjami wykorzystującymi AI do tzw. „rozbierania zdjęć" (z ang. *nudifying apps*). Stają się one coraz bardziej dostępne, zaawansowane, trudne do kontrolowania i stwarzają istotne problemy społeczne. Tego typu narzędzia umożliwiają generowanie realistycznych zdjęć przedstawiających osoby w nagiej formie. Ich bazę stanowią zazwyczaj neutralne fotografie, takie jak zdjęcia w mediach społecznościowych. AI umożliwia syntezę zdjęć bazowych, a następnie wygenerowanie duplikatu, na którym osoba widniejąca na zdjęciu jest pozbawiona ubrań[20].

Takie praktyki prowadzą do poważnych zagrożeń, których ofiarami padają często dzieci i młodzież i które powinny być kwalifikowane jako przejaw cyberprzemocy (cyberbullyingu). Zmanipulowane fotografie mogą być wykorzystywane do nękania i publicznego upokarzania ofiar, co prowadzi do daleko idących szkód psychicznych. Także w tym wypadku istotnym aspektem jest płeć pokrzywdzonych. Wskazuje się, iż nieproporcjonalnie często ofiarami ataków padają dziewczynki, co uwypukla problem dyskryminacji i uprzedmiotawiania kobiet. W Korei Południowej odnotowano kilkaset przypadków NSII w formie deep fakes wymierzonych w dziewczynki. Twórcy stworzyli ekosystem prześladowań, wymieniając się treściami za pośrednictwem komunikatorów[21]. Z pierwszych

18 Zob. Okolie C., op. cit.; Mania K. (2024). *Legal protection of revenge and deepfake porn victims in the European Union: findings from a comparative legal study.* „Trauma, Violence, & Abuse". Nr 25; Rigotti C., McGlynn C., *op. cit.*

19 Craft K. (2024). *Trolls Used Her Face to Make Fake Porn. There Was Nothing She Could Do.* https://www.nytimes.com/2024/07/31/magazine/sabrina-javellana-florida-politics-ai-porn.html.

20 Internet Matters (2024). *The new face of digital abuse: Children's experiences of nude deepfakes.* Internet Matters: Londyn; Gibson C. i in. (2024). *Analyzing the AI Nudification Application Ecosystem.* https://arxiv.org/pdf/2411.09751.

21 Mackenzie J.,Choi L. (2024). *Inside the Deepfake Porn Crisis Engulfing Korean Schools.* https://www.bbc.com/news/articles/cpdlpj9zn9go.

badań wynika, iż część ofiar na skutek przemocy decydowała się na istotne ograniczenie swojej aktywności społecznej[22].

Tego typu zjawiska przyczyniają się do normalizacji przemocy cyfrowej. Powszechna dostępność narzędzi do generowania tego typu szkodliwych treści promuje kulturę prześladowań, naruszania godności czy intymności. Co więcej, generowanie i udostępnianie publicznie materiałów prezentujących małoletnich może być wykorzystywane w produkcji treści pedofilskich.

Wygenerowane nagie zdjęcia mogą być wykorzystywane także do szantażu oraz realizacji scenariuszy tzw. „sextortion", w których sprawcy posługują się materiałami o charakterze kompromitującym uzyskanymi bez zgody ofiary w celu wymuszenia określonych korzyści, takich jak pieniądze, kolejne materiały czy przysługi seksualne[23]. W przypadku NSII w formie deep fakes sprawcy mogą tworzyć fałszywe treści, by zwiększyć wiarygodność swoich gróźb i wywrzeć większą presję na poszkodowaną osobę.

Istotnym problemem jest brak jednoznacznych przepisów karnych penalizujących tworzenie i udostępnianie NSII w formie deep fakes, co odnosi się także do polskiego Kodeksu karnego. Podobne trudności doprowadziły do rozpoczęcia prac nad kryminalizacją NSII w formie deep fakes w wielu jurysdykcjach (w tym w Wielkiej Brytanii i kilkunastu stanach USA), jak również pogłębionych analiz dotyczących rozszerzania interpretacji przepisów prawa karnego i cywilnego chroniących m.in. wizerunek i godność. W Wielkiej Brytanii, na mocy Online Safety Act utworzono nowe przestępstwo niekonsensualnego udostępniania cudzego wizerunku w kontekście seksualnym oraz przestępstwa kwalifikowane uwzględniające m.in. zamiar spowodowania cierpienia lub upokorzenia ofiary. Wprowadzenie przepisów stworzyło także silną podstawę prawną dla żądań wycofania nielegalnych treści z platform cyfrowych[24]. Na etapie prac podnoszono problem karalności za samo tworzenie NSII w formie deep fakes „na własny użytek" (bez intencji dystrybucji materiałów). Wskazuje się, że także tego typu sytuacje powinny być penalizowane, gdyż naruszają wizerunek i godność ofiar, a w przypadku wiedzy ofiary o posiadaniu przez sprawcę tego typu materiałów mogą wywoływać negatywne skutki o charakterze psychologicznym[25]. Dlatego też w styczniu 2025 r. rząd Wielkiej Brytanii podjął prace nad nowelizacją Online Safety Act w celu penalizacji tworzenia i posiadania NSII w formie deep fakes, w czym należy upatrywać podobieństw z wymogami unijnej Dyrektywy w sprawie zwalczania przemocy wobec kobiet i przemocy domowej.

W debacie prawnokarnej należy uwzględnić argumenty dotyczące praktycznego braku możliwości egzekwowania sankcji za tworzenie NSII w formie deep fakes na własny użytek, jednak nie powinny być pretekstem do zignorowania problemu na gruncie prawnym[26]. Na marginesie tych rozważań należy odnotować istotną analogię z posiadaniem treści prezentujących wykorzystywanie seksualne małoletnich (np. art. 202 § 4a), gdzie potencjalne trudności w praktycznym egzekwowaniu przepisu nie są przeciwskazaniem do kodeksowej penalizacji.

Rozwiązania wprowadzane w poszczególnych stanach USA różnią się od siebie, co prowadzi do ich

22 *Ibidem.*

23 FBI Public Service Announcement (2023). *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes.* Alert Number I-060523-PSA.

24 McGlynn C., Woods L., Antoniou A. (2024). *Pornography, the Online Safety Act 2023 and the need for further reform.* „Journal of Media Law".

25 Hörnle J. (2025). *Deepfakes and the Law: Why Britain needs stronger protections against technology-facilitated abuse.* https://www.qmul.ac.uk/media/news/2025/humanities-and-social-sciences/hss/deepfakes-and-the-law-why-britain-needs-stronger-protections-against-technology-facilitated-abuse.html.

26 *Ibidem;* McGlynn C., Woods L., *op. cit.*

fragmentacji. Nadzieję na unifikację i stworzenie jednolitych ram prawnych daje ponadpartyjna inicjatywa „Take It Down Act" procedowana przez Kongres, na mocy której tworzenie i udostępnianie NSII, także w formie deep fakes, miałoby być penalizowane, a platformy cyfrowe byłyby zobligowane do podjęcia odpowiednich kroków w celu zablokowania nielegalnych treści w ciągu 48 godzin od zgłoszenia[27].

Kraje Unii Europejskiej podejmują działania w zakresie stworzenia ram bezpiecznego rozwoju AI — zarówno na poziomie legislacji unijnej, jak i regulacji krajowych. Unijny Akt o Sztucznej Inteligencji (tzw. AI Act) definiuje deep fakes oraz przewiduje system zabezpieczeń przeciwko nim w postaci reguł transparentności. Są one jednak obliczone na przeciwdziałanie przede wszystkim dezinformacji politycznej.

W tym kontekście AI Act pomija problem wykorzystywania deep fakes do tworzenia NSII – reguły transparentności rozumiane przez pryzmat wyraźnego oznaczania syntetycznych treści nie są bowiem odpowiednim zabezpieczeniem[28]. Nawet jeśli syntetyczny charakter treści zostanie ujawniony wprost, taka klauzula nie eliminuje licznych dolegliwości niekorzystnie wpływających na psychikę i reputację ofiar[29]. Odniesienia do niekonsensualnych treści intymnych (jak również do materiałów przedstawiających wykorzystywanie seksualne dzieci) znalazły się natomiast w „Second Draft General-Purpose AI Code of Practice". Celem dokumentu jest m.in. wypracowanie standardów w zarządzaniu ryzykiem tworzonych systemów AI ogólnego przeznaczenia, co obejmuje również działania na rzecz ograniczenia wspomnianych szkodliwych treści poprzez monitoring gromadzenia i przetwarzania danych[30].

Luki w AI Act wypełnia przyjęta w 2024 r. Dyrektywa Parlamentu Europejskiego i Rady (UE) 2024/1385 z dnia 14 maja 2024 r. *w sprawie zwalczania przemocy wobec kobiet i przemocy domowej*, która obliguje państwa członkowskie do penalizacji tworzenia i udostępniania NSII w formie deep fakes (odniesienie do deep fakes można znaleźć bezpośrednio w motywie 19 dyrektywy). Transpozycja dyrektywy do prawa krajowego powinna zakończyć się do 14 czerwca 2027 r., co w wielu krajach UE wymagać będzie bezpośrednich zmian w prawie karnym.

Dyrektywa w sprawie zwalczania przemocy wobec kobiet i przemocy domowej odnosi się do konieczności penalizowania tworzenia i udostępniania NSII w formie deep fakes wprost: *„Publiczne udostępnianie za pomocą ICT obrazów, nagrań wideo lub podobnych treści przedstawiających czynności o wyraźnie seksualnym charakterze lub intymne części ciała danej osoby bez jej zgody może być bardzo szkodliwe dla ofiar [...]. Odnośne przestępstwo określone w niniejszej dyrektywie powinno obejmować wszystkie rodzaje takich treści, takie jak obrazy, zdjęcia i nagrania wideo, w tym obrazy oraz nagrania audio i wideo nacechowane seksualnie. Powinno ono odnosić się do sytuacji, w których publiczne udostępnianie treści za pomocą ICT odbywa się bez zgody ofiary, niezależnie*

27 Zob. m.in. *A Bill To require covered platforms to remove nonconsensual intimate visual depictions, and for other purposes*. https://files.constantcontact.com/d6c983c4801/af186e5b-7c99-4cb8-ac70-351913c62003.pdf?rdr=true. American O. (2025). *Pfluger, colleagues reintroduce TAKE IT DOWN* Act. https://www.oaoa.com/local-news/pfluger-colleagues-reintroduce-take-it-down-act.

28 Zob. m.in. Toparlak R. T. (2022). *Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse.* SciencesPo Law School The 10th Graduate Conference: Paryż; Centre for Digital Governance (2022). *The false promise of transparent deep fakes: How transparency obligations in the draft AI Act fail to deal with the threat of disinformation and image-based sexual abuse.* Hertie School: Berlin; Grady P. (2023). *EU Proposals Will Fail to Curb Nonconsensual Deepfake Porn.* https://datainnovation.org/2023/01/eu-proposals-will-fail-to-curb-nonconsensual-deepfake-porn.

29 Zob. m.in. Łabuz M. (2024). *Deep fakes and the Artificial Intelligence Act—An important signal or a missed opportunity?*. „Policy & Internet". Nr. 16(4).

30 Komisja Europejska (2024). *Second Draft of the General-Purpose AI Code of Practice published, written by independent experts.* https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts.

od tego, czy ofiara wyraziła zgodę na wytworzenie takich treści bądź przesłała je konkretnej osobie. Przestępstwo to powinno również obejmować wytwarzanie treści bez zgody lub manipulowanie nimi lub ich zmienianie, na przykład przez edytowanie obrazów, między innymi z wykorzystaniem sztucznej inteligencji, które to treści sprawiają wrażenie, że osoba uczestniczy w czynnościach seksualnych, o ile treści te są następnie udostępniane publicznie za pośrednictwem ICT bez zgody tej osoby. Takie wytwarzanie, manipulowanie lub zmienianie powinno obejmować wytwarzanie treści «deepfake», w których prezentuje się osoby, przedmioty, miejsca lub inne podmioty bądź wydarzenia wyraźnie podobne do istniejących naprawdę, i które przedstawiają czynności seksualne danej osoby i błędnie wydają się innym osobom autentyczne lub prawdziwe. W interesie skutecznej ochrony ofiar takich czynów należy również uwzględnić groźby dopuszczenia się takich czynów"[31].

Zaproponowane w Dyrektywie rozwiązania obejmują stworzenie kompleksowego systemu przeciwdziałania przemocy względem kobiet, w tym w Internecie. Wśród nich wymienić należy konieczność zapewnienia odpowiedniej pomocy, w tym wsparcia psychologicznego, ofiarom przemocy, współpracę z platformami cyfrowymi w celu szybszej identyfikacji i moderacji treści abuzywnych, czy działania w sferze edukacji i zwiększania świadomości[32].

W związku z koniecznością zajęcia się problemem wykorzystywania deep fakes do tworzenia niekonsensualnych treści o charakterze seksualnym, w tym zabezpieczenia praw ofiar oraz istotnego zwiększenia świadomości po stronie tworzących prawo i społeczeństwa, niezbędne są zmiany w Kodeksie karnym w kwalifikowaniu tworzenia i udostępniania takich treści. Wprowadzenie jednoznacznych przepisów penalizujących wprost tworzenie i udostępnianie NSII w formie deep fakes jest szansą na stworzenie silniejszego systemu ochrony ofiar oraz skuteczną transpozycję wskazanej Dyrektywy.

Obecnie w polskim prawie karnym nie istnieją przepisy, które wprost penalizowałyby tworzenie i udostępnianie materiałów o charakterze NSII w formie deep fakes. Wobec braku szczegółowych regulacji, czyny te mogłyby być kwalifikowane subsydiarnie na podstawie istniejących przepisów, takich jak art. 191a k.k., który penalizuje rejestrowanie i rozpowszechnianie wizerunku nagiej osoby bez jej zgody, czy art. 212 i 216 k.k. dotyczących zniesławienia i zniewagi[33]. Art. 190a k.k., dotyczący uporczywego nękania (stalkingu), mógłby być używany do ścigania działań związanych z publikowaniem NSII w formie deep fakes w celu zastraszenia ofiary, ale nie obejmowałby sytuacji, w których materiały te są tworzone i dystrybuowane bez takiego zamiaru. Wskazane przepisy nie mają ugruntowanej interpretacji odnoszącej się do problematyki NSII w formie deep fakes[34]. Co więcej, w literaturze podnosi się, iż nawet art. 191a k.k. może nie mieć zastosowania w przypadku kreacji hybrydowych, tzn. gdy na wizerunek aktorki porno nałożono twarz pokrzywdzonej. Wówczas bowiem wizerunek nagiej osoby nie należy do osoby pokrzywdzonej, lecz dochodzi do powielenia wizerunku zarejestrowanego (z dużym prawdopodobieństwem) na skutek konsensualnej czynności[35]. W przypadku przestęp-

31 Dyrektywa Parlamentu Europejskiego i Rady (UE) 2024/1385 z dnia 14 maja 2024 r. *w sprawie zwalczania przemocy wobec kobiet i przemocy domowej.*

32 *Ibidem;* Parlament Europejski (2024). *Cyberprzemoc wobec kobiet: czym jest i jak jej zapobiegać?* https://www.europarl.europa.eu/topics/pl/article/20241205STO25880/cyberprzemoc-wobec-kobiet-czym-jest-i-jak-jej-zapobiegac.

33 Ziobroń A., *op cit.*; Vera G. G., *op. cit.*.

34 Zob. Ziobroń A., *op. cit.*; Sewastianowicz M. (2023). *Deepfake - ofiara realistycznej przeróbki może mieć problem z dochodzeniem swoich praw.* https://www.prawo.pl/prawo/deepfake-a-prawo-karne-film-porno,520138.html.; Kupis M., Łaguna Ł., (2023). *Czy deepfake jest w Polsce legalny?.* https://law4tech.pl/czy-deepfake-jest-w-polsce-legalny.

35 Ziobroń A., *op. cit.*

czości seksualnej zasadne wydaje się zrewidowanie przepisów dotyczących ochrony wizerunku bądź też przestępstw na tle seksualnym, które w bardzo słabym stopniu rejestrują zjawisko AI jako samoistny problem wymagający reakcji prawnej.

Art. 191a. § 1. Kto utrwala wizerunek nagiej osoby lub osoby w trakcie czynności seksualnej, używając w tym celu wobec niej przemocy, groźby bezprawnej lub podstępu, albo wizerunek nagiej osoby lub osoby w trakcie czynności seksualnej bez jej zgody rozpowszechnia, podlega karze pozbawienia wolności od 3 miesięcy do lat 5. § 2. Ściganie następuje na wniosek pokrzywdzonego[36].

W kontekście polskiego Kodeksu karnego należałoby zatem przemyśleć potrzebę penalizacji tworzenia i rozpowszechniania NSII w formie deep fakes w sposób konkretny, co z kolei wymagałoby wprowadzenia jednoznacznie brzmiącego przepisu i czyniłoby zadość wymogom zawartym w Dyrektywie w sprawie zwalczania przemocy wobec kobiet i przemocy domowej. Zasadne wydaje się doprecyzowanie art. 191a k.k. i rozszerzenie go o dodatkowy paragraf, który jasno wskazywałby, iż penalizowane jest także tworzenie i udostępnianie wytworzonego lub przetworzonego wizerunku identyfikowalnej osoby, w tym z wykorzystaniem narzędzi AI lub w formie deep fake. Odniesienie do „identyfikowalnej" osoby miałoby na celu wykluczenie sytuacji penalizowania tworzenia całkowicie syntetycznej pornografii przedstawiającej osoby dorosłe, gdy nie doszło do naruszenia praw osób trzecich. Takie czynności, co do zasady, nie powinny bowiem podlegać sankcji karnej.

Na marginesie warto odnotować, iż dodatkowe zabezpieczenia przed NSII w formie deep fakes mogą pojawić się na gruncie prawa cywilnego. W obecnym stanie prawnym ofiary mogą próbować dochodzić swoich praw na podstawie przepisów o naruszeniu dóbr osobistych, w tym prawa do wizerunku, dobrego imienia czy prywatności. Wobec braku jednoznacznej podstawy prawnej i rozwoju technologii pozwalającej na daleko idące manipulacje obrazem i dźwiękiem zasadne byłoby bardziej szczegółowe uregulowanie wizerunku i zakresu jego ochrony oraz posługiwanie się konkretną definicją deep fakes, dla której wzór powinny stanowić regulacje UE. Pozwoliłoby to również na skrócenie czasochłonnych procesów sądowych, ograniczając kosztowne i dodatkowo traumatyzujące ekspertyzy prawne, a jednocześnie sprzyjać szybszemu usuwaniu nielegalnych treści przez platformy cyfrowe. Uproszczenie procedury dochodzenia praw ofiar NSII w formie deep fakes stanowiłoby zatem element przeciwdziałania wtórnej wiktymizacji i czynnik wspierający postulowane w dalszej części opracowania działania o charakterze pozaprawnym. Uwzględnienie problemu wywierania większej presji na działania podejmowane przez platformy cyfrowe, co jest postulowane m.in. w Wielkiej Brytanii czy USA, byłoby z kolei elementem budowania ekosystemu ochronnego, dla którego podstawy stanowić może unijny Akt o Usługach Cyfrowych (DSA) obligujący platformy cyfrowe do mitygowania ryzyk systemowych i odpowiedniej moderacji treści, w tym usuwania treści abuzywnych[37].

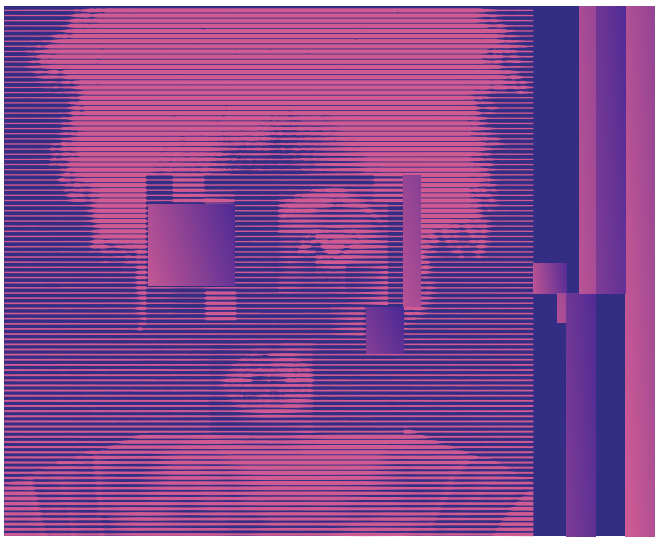
Proponowane regulacje powinny uwzględniać ochronę wolności wypowiedzi. Ustawodawca musi mieć na uwadze, aby nowe przepisy nie prowadziły do pogłębiania problemu nadmiernej moderacji treści dostępnych w sieci, a jednocześnie uwzględniać konieczność szybkiego i efektywnego egzekwowania przepisów w przypadku

36 Ustawa z dnia 6 czerwca 1997 r. *Kodeks karny* (Dz. U. 1997 Nr 88 poz. 553)

37 Karaboga M. (2023). *Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags.* [w:] Jaki S., Steiger S. (ed.) *Digitale Hate Speech.* J.B. Metzler: Berlin, Heidelberg.

uzasadnionych żądań usunięcia treści naruszających prawa osób trzecich. W konsekwencji konieczne w ramach procesu jest również uwzględnienie efektywności już istniejących narzędzi wdrożonych przez platformy cyfrowe, takich jak mechanizmy zgłaszania i filtrowania oraz systemy flagowania stron zawierających tego rodzaju treści i wywieranie nacisku na ich egzekwowanie i zwiększanie tempa reakcji. Sytuacje takie jak zarejestrowane w 2024 r. na platformie X udostępnianie NSII w formie deep fakes prezentujących piosenkarkę Taylor Swift wzbudzają uzasadnione obawy – fałszywe materiały były dostępne przez kilkadziesiąt godzin, zostały wyświetlone ok. 50 milionów razy i dopiero zablokowanie możliwości wyszukiwania danych Swift umożliwiło ich moderację[38]. Z badań przeprowadzonych w 2024 r. wynika, iż czas oczekiwania na moderację treści naruszających politykę dot. treści prezentujących nagie osoby wynosił nawet do trzech tygodni[39]. Można mieć zatem uzasadnione obawy, iż system przewidziany w DSA w przypadku NSII w formie deep fakes nie działa skutecznie i wymaga dalszych prac.

Z punktu widzenia egzekwowania zaproponowanych przepisów karnych niezbędne jest zwrócenie uwagi na trudności ścigania przypadków cyberprzemocy. Skuteczność technik zapewniających anonimowość sprawcom przestępstw może przyczyniać się do ich bezkarności, co w czarnym scenariuszu mogłoby doprowadzić do faktycznej bezradności organów ścigania, dając ofiarom jedynie iluzoryczną ochronę. Nie jest to jednak argument świadczący przeciwko wprowadzaniu odpowiednich przepisów. Wprost przeciwnie, sformułowanie konkretnych przepisów to niezbędny element przeciwdziałania przemocy online na tle seksualnym oraz okazja do wysłania sygnału, iż tego typu czyny przestępcze będą ścigane. Należy założyć, iż konkretyzacja przepisów będzie działać odstraszająco na wybranych sprawców, w czym można upatrywać okazji do zmniejszenia liczby przypadków tworzenia i udostępniania NSII.



38 Saner E. (2024). *Inside the Taylor Swift Deepfake Scandal: 'It's Men Telling a Powerful Woman to Get Back in Her Box'.* https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box.

39 Qiwei L. i in. (2024). *Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes.* PrePrint: https://arxiv.org/pdf/2409.12138.

NSII w formie deep fakes – propozycje działań pozaprawnych

Prócz zmian na gruncie prawnym niezbędne jest także zwiększanie działań pomocowych, budowanie świadomości społecznej w zakresie zagrożeń, oraz podnoszenie problematyki przemocy i dyskryminacji na tle płciowym w wymiarze cyfrowym. Samo wprowadzenie przepisów penalizujących tworzenie lub upowszechnianie NSII w formie deep fakes nie rozwiązuje bowiem problemu cierpienia ofiar, które może być łagodzone poprzez stosowanie innych środków naprawczych.

Istotnym elementem są działania pozaprawne, które powinny zostać podjęte przez instytucje publiczne, co może uwzględniać doświadczenia innych krajów i obejmować: tworzenie telefonów zaufania, zapewnianie profesjonalnej pomocy prawnej i psychologicznej ofiarom NSII w formie deep fakes, wdrażanie programów edukacyjnych i podnoszenie świadomości, współpracę z platformami cyfrowymi w celu efektywnej moderacji treści w odpowiedzi na zgłoszenia naruszeń, jak również inwestycje w technologie umożliwiające wykrywanie nielegalnego wykorzystania wizerunku, czemu sprzyjać może współpraca z sektorem prywatnym. Istotną rolę w tych procesach powinny odgrywać Ministerstw Cyfryzacji, Zdrowia, Rodziny, Pracy i Polityki Społecznej, Edukacji Narodowej, ale także Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK) jako jedna z kluczowych instytucji zajmujących się bezpieczeństwem w cyberprzestrzeni, która już ma bogate doświadczenia oraz istotny poziom ekspertyzy w działaniach wzmacniających społeczną świadomość i odporność.

Istotnym aspektem mogą być działania podejmowane na poziomie systemu edukacji. W tym kontekście warto uwzględnić przeciwdziałanie szkodliwym formom wykorzystania AI, co powinno obejmować NSII w formie deep fakes i różne formy tego zjawiska, w tym wykorzystywanie aplikacji rozbierających. Istotnym problemem jest brak świadomości szkodliwości tego typu działań po stronie dzieci, młodzieży, ale i osób dorosłych, jak również trywializacja niektórych szkodliwych zachowań. Także w tym wypadku konieczne jest monitorowanie trendów i reagowanie na negatywne zjawiska zaobserwowane w wielu krajach, co powinno być przedmiotem zainteresowania wymiaru sprawiedliwości oraz właściwych resortów i NASK. Na poziomie wymiaru sprawiedliwości konieczne są dodatkowe szkolenia dla personelu, w tym sędziów, w celu zwiększenia kompetencji w obszarze nowoczesnych technologii oraz uwrażliwienia na problematykę NSII.

Programy edukacyjne powinny obejmować zwiększanie świadomości społecznej w odniesieniu do zagrożeń i ich konsekwencji oraz podnoszenie problematyki cyberprzemocy z wykorzystaniem nowoczesnych technologii. Powinny im towarzyszyć działania pomocowe, w tym psychologiczne, oferowane również w samych placówkach edukacyjnych, co z kolei wymaga odpowiednich kompetencji personelu. Ich rozwijanie powinno uwzględniać szkolenia dla nauczycieli i szkolnych psychologów. W wymiarze holistycznym tego typu działania są elementem budowania społecznej odporności i wzmacniania demokratycznych wartości.

CSAM w formie deep fakes – propozycje zmian w prawie karnym

Deep fakes od wielu lat uznawane są za istotne zagrożenie dla przestrzeni informacyjnej w kontekście stopniowego zacierania granicy między treściami prawdziwymi i syntetycznymi. Wskazuje się także na ich potencjał do fałszowania dowodów w procesach sądowych[40]. Wobec znacznego upowszechnienia i uproszczenia technologii tworzenia syntetycznych mediów, deep fakes są także coraz częściej wykorzystywane do generowania tzw. Child Sexual Abuse Materials (CSAM), które w polskim porządku prawnym są opisywane jako treści pornograficzne z udziałem małoletniego (art. 202 par. 3 k.k.).

W odniesieniu do treści prezentujących wykorzystywanie seksualne małoletnich polski Kodeks karny posługuje się określeniem „treści pornograficzne z udziałem małoletniego". Ze względów semantycznych bardziej wskazane byłoby operowanie określeniem Child Sexual Abuse Materials (CSAM), które jest stosowane zamiennie z określeniem kodeksowym w dalszej części opracowania, i uwypukla aspekt przemocy i wykorzystywania seksualnego.

Polski Kodeks karny nie odnosi się wprost do deep fakes w przypadku CSAM, jednak w art. 202 par. 4b penalizuje produkowanie, rozpowszechnianie, prezentowanie, przechowywanie i posiadanie treści pornograficznych przedstawiających wytworzony albo przetworzony wizerunek małoletniego uczestniczącego w czynności seksualnej[41]. Odniesienie do „wytworzonego lub przetworzonego wizerunku" powinno być współcześnie interpretowane także w kontekście mediów syntetycznych. W tym względzie polski Kodeks karny pozwala na dostosowanie przepisów do rozwoju nowoczesnych technologii i kwalifikowanie nowych form popełniania przestępstw. Choć przepis ten był tworzony w innych realiach historycznych i technologicznych, powinien być interpretowany adekwatnie do pojawienia się nowych możliwości wytwarzania i przetwarzania wizerunku małoletnich, w tym z wykorzystaniem deep fakes[42].

Art. 202. § 4b. Kto produkuje, rozpowszechnia, prezentuje, przechowuje lub posiada treści pornograficzne przedstawiające wytworzony albo przetworzony wizerunek małoletniego uczestniczącego w czynności seksualnej podlega karze pozbawienia wolności do lat 3[43].

Ze względu na dobór sankcji art. 202 par. 4b k.k. nie jest jednak w pełni adekwatny wobec zwiększającej się liczby mediów syntetycznych prezentujących CSAM oraz ich postępującej jakości skutkującej faktyczną nieodróżnialnością materiałów prawdziwych

40 Delfino R. (2023). *The Deepfake Defense—Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers.* „Ohio State Law Journal". Nr 84; Llorente R. V. (2024). *Deepfakes in the Dock: Preparing International Justice for Generative AI.* „The SciTech Lawyer". Nr 20(2).

41 Ustawa z dnia 6 czerwca 1997 r. *Kodeks karny* (Dz. U. 1997 Nr 88 poz. 553).

42 Zob. m.in. Niedbała M. (2023). *Dylemat odpowiedzialności karnej z tytułu generowania pornografii przy wykorzystaniu sztucznej inteligencji.* „Krytyka Prawa". Nr 15(4); Staciwa K. (2023). *Wykorzystywanie seksualne dzieci w cyberprzestrzeni.* „Dziecko Krzywdzone. Teoria, badania, praktyka". Nr 22(3); Hypś S. (2024). Teza V.3, [w:] *Kodeks karny. Komentarz* (red. A. Grześkowiak, K. Wiak). C.H. Beck: Warszawa.

43 Ustawa z dnia 6 czerwca 1997 r. *Kodeks karny* (Dz. U. 1997 Nr 88 poz. 553).

i wygenerowanych przy użyciu AI. W wybranych krajach (m.in. Kanada, Korea Południowa, Wielka Brytania), gdzie zapadły pierwsze precedensowe wyroki w sprawach dotyczących tworzenia i udostępniania CSAM w formie deep fakes, orzecznictwo zwróciło uwagę na szereg dodatkowych problemów wiążących się z wykorzystywaniem technologii do multiplikowania szkodliwych treści[44].

Wspomniana nieodróżnialność treści prawdziwych (realnych nagrań przedstawiających prawdziwe dzieci) oraz syntetycznych (stworzonych przy wykorzystaniu AI) znacząco utrudnia pracę organów ścigania i biegłych, doprowadzając do istotnego zwiększenia liczby materiałów, jakie należy przeanalizować, co z kolei może skutkować nadmiernym obciążeniem ekspertów, angażowaniem sił i środków, a w konsekwencji osłabieniem ochrony ofiar pedofilów[45]. Odnotowano także, iż multiplikacja treści przyczynia się do zwiększania społecznej szkodliwości, w tym poprzez normalizację występowania treści pornograficznych z udziałem małoletnich[46]. Precedensowe wyroki pokazały trudności z odpowiednim kwalifikowaniem i penalizowaniem tego typu czynów.

W związku z narastającym problemem niezbędne jest podjęcie tematu penalizacji tworzenia i udostępniania CSAM na gruncie polskiego prawa karnego, w tym nieadekwatności sankcji aktualnych przepisów dotyczących treści pornograficznych przedstawiających wytworzony albo przetworzony wizerunek małoletniego uczestniczącego w czynności seksualnej (art. 202. § 4b k.k.). Zasadne wydaje się dokonanie rozróżnienia między kilkoma typami czynów zabronionych obejmujących następujące zmienne: 1) wykorzystanie seksualne prawdziwego dziecka (czy treść została zarejestrowana z udziałem prawdziwego dziecka, czy też została wygenerowana przez AI bez udziału prawdziwego dziecka); 2) wykorzystanie wizerunku prawdziwego dziecka (czy treść została wygenerowana z wykorzystaniem wizerunku prawdziwego dziecka, czy też została wygenerowana w całości przez AI bez identyfikowalnego wizerunku prawdziwego dziecka); 3) aktywność sprawcy (czy materiały zostały stworzone na własny użytek, czy też doszło do ich udostępnienia). Każdy z wymienionych wyżej aspektów powinien wpłynąć na ocenę stopnia społecznej szkodliwości czynów, a przez to na wymiar sankcji.

Na poziomie tworzenia treści pornograficznych z udziałem małoletniego najwyższą sankcją powinny być zagrożone czyny, w których dochodzi do skrzywdzenia prawdziwego dziecka. Zasadne wydaje się tutaj odróżnienie sytuacji, w których do stworzenia treści pornograficznych z udziałem małoletniego wykorzystana została synteza AI. Dobrem chronionym jest w tym wypadku bowiem dobro indywidualne dziecka.

Na poziomie rozpowszechniania, odbioru i posiadania CSAM stopień społecznej szkodliwości jest z jednej strony wyznaczany dobrem indywidualnym dziecka, jeśli posłużono się jego prawdziwym lub zmodyfikowanym wizerunkiem (np. przy wtórnej wiktymizacji, gdy wcześniej zarejestrowany materiał prezentujący prawdziwe dziecko jest wykorzystywany do dalszej syntezy), z drugiej multiplikacją treści uznawanych obiektywnie za szkodliwe. W konsekwencji, na poziomie rozpowszechniania, odbioru i posiadania CSAM, ze względu na nieodróżnialność treści prawdziwych (w których zarejestrowano rzeczywiste wykorzystanie seksualne małoletniego) i syntetycznych (wygenerowanych przez AI) sankcje nie powinny się od siebie różnić.

44 Zob. m.in. wyrok Sądu Prowincjonalnego w Quebec z 14 kwietnia 2023 r. w sprawie przeciwko S. Larouche.

45 Internet Watch Foundation, *op. cit.*; Harwell D., *AI-generated child sex images spawn new nightmare for the web*, https://www.washingtonpost.com; NCA (2024). *Director General Graeme Biggar launches National Strategic Assessment,* https://www.nationalcrimeagency.gov.uk.

46 *Ibidem*.

Z punktu widzenia odbiorcy takich treści obiektywny brak możliwości rozróżnienia treści prawdziwych od syntetycznych (wytworzonych lub przetworzonych) powinien prowadzić do braku różnicowania naganności inkryminowanych czynności.

Zalecana jest zatem modyfikacja systemu kar (w aktualnym porządku prawnym zwiększenie maksymalnej sankcji wynikającej z art. 202 § 4b k.k.) i rozróżnienie odpowiedzialności w zależności od genezy wytworzenia treści oraz roli sprawcy – twórcy, dystrybutora, odbiorcy czy posiadacza.

Rozszerzenie zakresu ochrony prawnej na treści syntetyczne wymaga dodatkowo uwzględnienia szerszego kontekstu społecznego, w tym wpływu na percepcję seksualizacji dzieci i utrudniania pracy organów ścigania. W konsekwencji dystrybucja CSAM w formie deep fakes, które są nieodróżnialne od prawdziwych obrazów z udziałem dzieci, powinna być traktowana na równi z rzeczywistą CSAM, niezależnie od braku fizycznego udziału dziecka w procesie tworzenia materiału. Rozpiętość sankcji dawałoby z kolei możliwość wymierzenia wyższej kary, gdy wykorzystanie wizerunku prawdziwego dziecka podnosiłoby stopień społecznej szkodliwości (w przypadku świadomej dystrybucji takiego materiału).

Wyzwanie w zakresie efektywnego przeciwdziałania wyżej wskazanym zagrożeniom stanowi opracowanie przepisów, które pozwolą na skuteczną walkę z nasilającym się problemem, przy jednoczesnym uniknięciu nieproporcjonalnych naruszeń prywatności użytkowników Internetu. W toku prac nad regulacjami dotyczącymi CSAM na poziomie UE należy zapewnić nienaruszalność zasady szyfrowania prywatnych wiadomości, co jednak nie wpływa istotnie na problem wysokości sankcji i rozróżnienia typów czynów zabronionych omówionych powyżej na gruncie polskiego prawa karnego.

Na marginesie tych rozważań należy wspomnieć o rozpoczęciu prac legislacyjnych w Wielkiej Brytanii (luty 2025 r.), których celem jest zdelegalizowanie posiadania narzędzi umożliwiających tworzenie syntetycznych CSAM (zagrożone karą więzienia do pięciu lat) oraz posiadania instrukcji, w jaki sposób to zrobić (zagrożone karą więzienia do trzech lat)[47]. Wielka Brytania byłaby pierwszym krajem z tego typu regulacjami. Ich ocena będzie jednak zależna od konkretnych sformułowań - o ile penalizowanie posiadania „samouczków" nie powinno rodzić większych kontrowersji, o tyle problematyczne może być stworzenie przepisu zakazującego posiadania konkretnego oprogramowania. Jak wskazują badania, do tworzenia CSAM w formie deep fakes wykorzystywane są także modele AI open source, w tym popularny Stable Diffusion[48].

47 Internet Watch Foundation (2025). *New AI child sexual abuse laws announced following IWF campaign*. https://www.iwf.org.uk/news-media/news/new-ai-child-sexual-abuse-laws-announced-following-iwf-campaign.

48 Crawford A., Smith T. (2023). *Illegal trade in AI child sex abuse images exposed*. https://www.bbc.com/news/uk-65932372; Office of Public Affairs US Department of Justice (2023). *Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct*. https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged.

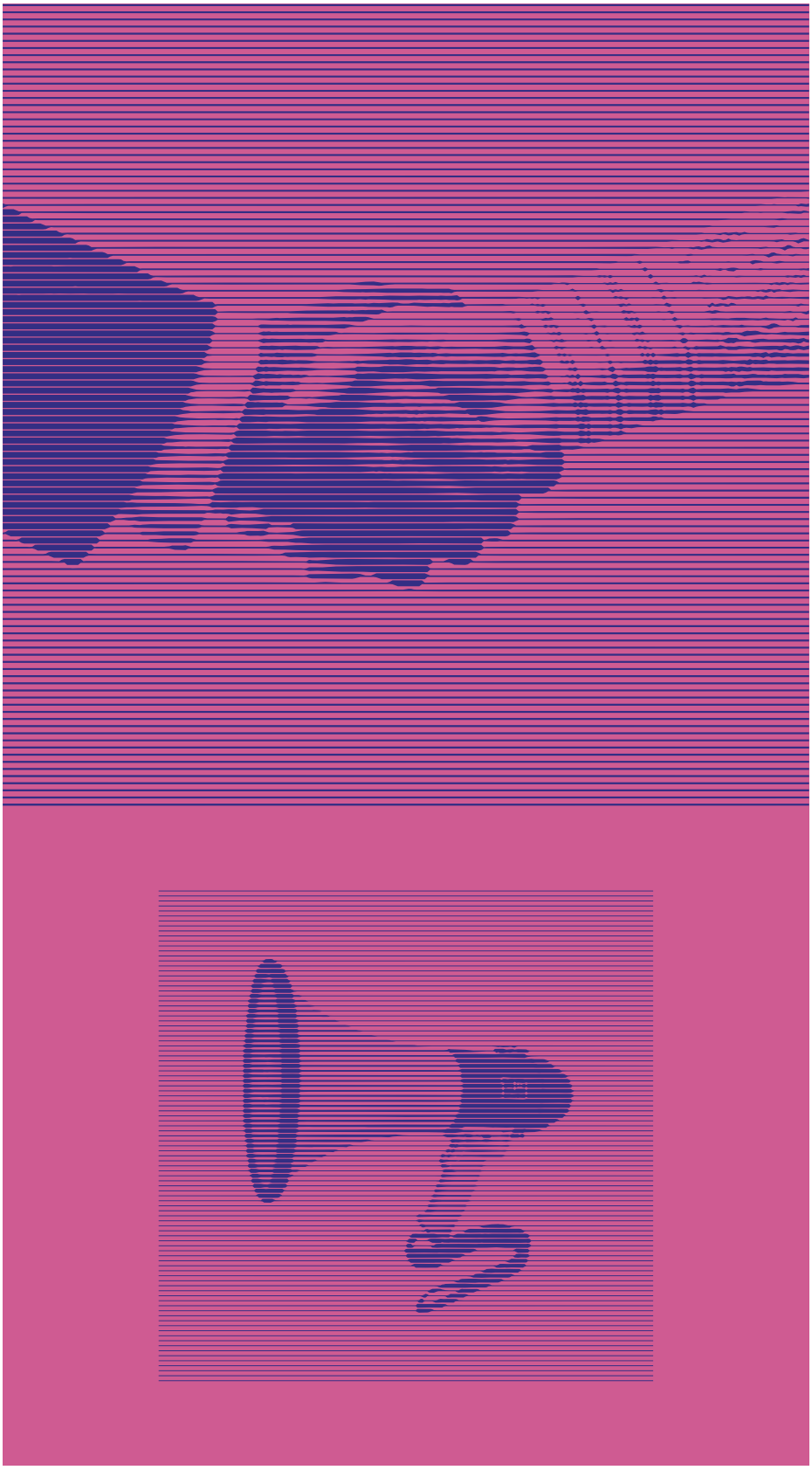
CSAM w formie deep fakes – propozycje działań pozaprawnych

Adekwatne sankcjonowanie tworzenia, dystrybucji, odbioru i posiadania CSAM w formie deep fakes nie wyczerpuje innych problemów związanych z upowszechnieniem tego typu treści w przestrzeni cyfrowej. Jak wskazano, szczególnie niepokojące jest zjawisko multiplikacji syntetycznego CSAM skutkujące nadmiernym obciążeniem organów ścigania i biegłych. W konsekwencji może to prowadzić do braku możliwości prowadzenia efektywnych śledztw i rozpoznawania prawdziwych (nie-syntetycznych) ofiar CSAM.

W tym kontekście niezbędne jest podejmowanie działań na rzecz wzmocnienia odporności systemu prawnego poprzez inwestowanie w kapitał ludzki (szkolenia, rozwijanie kompetencji) oraz nowoczesne technologie umożliwiające przynajmniej częściowe odciążenie organów ścigania i biegłych. W pierwszym z wymienionych elementów budowania odporności niezbędne są dodatkowe szkolenia i konsekwentne podnoszenie kompetencji w zakresie wykrywania treści syntetycznych. Tego typu działania powinny być uzupełniane inwestycjami w oprogramowanie pozwalające na wykrywanie syntetycznych mediów, a co za tym idzie także szkoleń uwzględniających rozwój technologii. Współpraca z sektorem prywatnym może zapewnić odpowiedni transfer wiedzy. Mimo iż oprogramowanie do wykrywania syntetycznych mediów nie może być na ten moment uznawane za w pełni wiarygodne, a jego efektywność jest oceniana poniżej 100%, może być czynnikiem wspomagającym pracę organów ścigania i biegłych, a przez to pozwalać na ograniczenie czasu niezbędnego do przeprowadzenia ekspertyzy.

W odniesieniu do wdrażania nowoczesnych technologii w szeroko pojętym wymiarze sprawiedliwości istotne jest uwzględnienie współpracy z sektorem prywatnym, szczególnie w przypadku wykorzystania modeli przeznaczonych wprost do wykrywania syntetycznego CSAM. Identyfikowanie podmiotów posiadających ekspertyzę w tej dziedzinie i utrzymywanie odpowiednich relacji powinno stać się jednym z obowiązków aparatu administracji publicznej.

Na poziomie społecznym istotne jest uwzględnienie konieczności zwiększania świadomości, przede wszystkim poprzez przeciwdziałanie nadmiernemu upublicznianiu wizerunków dzieci w mediach społecznościowych przez rodziców. Zjawisko to nazywane „sharentingiem" może bowiem dostarczać danych wejściowych do trenowania modeli AI przewidzianych do tworzenia treści o charakterze CSAM bądź też stać się bezpośrednią bazą dla syntezy. To z kolei może prowadzić do wiktymizacji dzieci, których wizerunek będzie wykorzystywany do stworzenia CSAM w formie deep fakes. Zalecane są zatem kampanie informacyjne, które jednocześnie wskazywać będą na istotny wzrost zagrożeń z zakresu cyberbezpieczeństwa związanych z możliwością wykorzystywania deep fakes do kradzieży wizerunku oraz głosu.



Rekomendacje

Ochrona ofiar NSII w formie deep fakes na gruncie prawa karnego

Wprowadzenie konkretnych przepisów wprost penalizujących tworzenie i rozpowszechnianie NSII w formie deep fakes. W tym kontekście zalecane jest przyspieszenie prac nad wdrożeniem przepisów wynikających z Dyrektywy Parlamentu Europejskiego i Rady UE 2024/1385 z dnia 14 maja 2024 r. *w sprawie zwalczania przemocy wobec kobiet i przemocy domowej.* Zasadne w tym względzie jest znowelizowanie art. 191a k.k. i wprowadzenie jednoznacznego odniesienia do treści generowanych przy użyciu AI. Ostateczny termin wdrażania Dyrektywy wyznaczony na czerwiec 2027 r. nie powinien stanowić punktu odniesienia dla prac na gruncie krajowym. Mając na uwadze pilność działań, wdrożenie Dyrektywy do polskiego porządku prawnego powinno nastąpić jak najszybciej.

Dodatkowa ochrona ofiar NSII w formie deep fakes

Zastosowanie środków cywilnych i administracyjnych w celu umożliwienia szybkiego usuwania z Internetu treści abuzywnych naruszających prawa osób trzecich. Wymaga to większego nadzoru nad egzekwowaniem prawa w odniesieniu do platform cyfrowych i uwzględnienia kontekstu NSII w formie deep fakes w ramach stosowania unijnego Aktu o Usługach Cyfrowych. Niezbędne jest również zapewnienie ofiarom kompleksowej pomocy psychologicznej i prawnej, np. poprzez wprowadzenie telefonu zaufania oraz bezpłatnych konsultacji prawnych dla pokrzywdzonych.

Edukacja i świadomość społeczna

Wdrożenie i rozszerzanie programów edukacyjnych skierowanych do dzieci, młodzieży i dorosłych, uwzględniających zagrożenia związane z deep fakes, cyberprzemocą i niewłaściwym wykorzystywaniem technologii AI. Edukacja na poziomie szkolnym (co wymaga zaangażowania MEN) powinna obejmować zagadnienia etycznego korzystania z technologii i podejmować m.in. temat szkodliwości aplikacji pozwalających na tworzenie „rozbieranych zdjęć". W tym kontekście istotne jest również monitorowanie i mapowanie przypadków wiktymizacji oraz realizowanie programów podnoszących świadomość (np. przez Ministerstwo Cyfryzacji, NASK).

Nowelizacja art. 202 § 4b k.k. i sankcjonowanie CSAM w formie deep fakes

Podniesienie maksymalnych sankcji w odniesieniu do rozpowszechniania CSAM w formie deep fakes ze względu na istotny postęp jakościowy syntezy AI i faktyczną nieodróżnialność treści wytworzonych lub przetworzonych w taki sposób, co można osiągnąć poprzez odpowiednią nowelizację art. 202 § 4b k.k. Alternatywnym rozwiązaniem jest stworzenie nowego przepisu umożliwiającego precyzyjne zdefiniowanie w przepisach CSAM w formie deep fakes, co ułatwi ich penalizację i eliminację luk interpretacyjnych.

Sankcje za treści syntetyczne i rzeczywiste

Uwzględnienie w przepisach karnych zróżnicowania sankcji na poziomie tworzenia CSAM w formie deep fakes w zależności od tego, czy ucierpiało prawdziwe dziecko (analiza genezy treści: czy wykorzystano rzeczywiste dziecko, czy stworzono treści w pełni syntetyczne). Sankcje za rozpowszechnianie syntetycznych CSAM nieodróżnialnych od prawdziwych treści powinny być równie wysokie, jak za rzeczywiste CSAM, aby uwzględnić ich społeczną szkodliwość oraz utrudnienia w ściganiu takich przestępstw. Zgodnie z zasadami postępowania karnego, końcowy wymiar kary zależeć będzie od interpretacji elementów składających się na czyn przestępny, co jest naturalną konsekwencją indywidualnego podejścia do konkretnych przypadków.

Wzmocnienie zdolności technologicznych organów ścigania

Inwestowanie w technologie umożliwiające skuteczne wykrywanie syntetycznych treści oraz szkolenia dla ekspertów i biegłych w zakresie identyfikacji i analizy CSAM w formie deep fakes. Zwiększaniu zdolności i kompetencji może towarzyszyć zacieśnienie współpracy międzynarodowej w zakresie ścigania przestępstw związanych z deep fakes, w tym międzynarodowa i międzyinstytucjonalna wymiana informacji oraz dobrych praktyk, jak również monitorowanie trendów i współpraca z sektorem prywatnym. Współpraca w wymiarze ponadnarodowym powinna obejmować również nacisk na platformy cyfrowe w zakresie wykrywania nielegalnych treści, monitoringu ich wolumenu, jak również moderacji treści i szybkiego reagowania na zgłoszenia.

Dwutorowe podejście do regulowania deep fakes

Oddzielenie problematyki NSII od szerszego dyskursu na temat deep fakes i ich regulowania. Zaproponowane działania koncentrują się wyłącznie na przeciwdziałaniu NSII i jego konsekwencjom. Dyskusja na temat penalizowania tworzenia i udostępniania NSII nie powinna obejmować słusznej, acz odmiennej, debaty na temat regulowania deep fakes jako takich, w szczególności o charakterze dezinformacji politycznej. Mieszanie tych dwóch sfer może bowiem negatywnie wpływać na gotowość do podjęcia tematu NSII przez decydentów politycznych i społeczeństwo i być pretekstem do oskarżeń o próby cenzurowania Internetu i ograniczania wolności słowa. Dodatkowo, zarówno w przypadku NSII jak i CSAM w formie deep fakes wprowadzenie konkretnych przepisów karnych będzie miało wydźwięk psychologiczny, przeciwdziałając deprecjacji szkodliwości opisywanych czynów i trywializacji zagrożeń.

Monitorowanie trendów legislacyjnych

Monitorowanie trendów legislacyjnych oraz precedensowych rozstrzygnięć sądowych. Szczególnie istotne jest obserwowanie aktywności legislacyjnej w Wielkiej Brytanii, która jest pionierem w zakresie ustawodawstwa przeciwdziałającego CSAM w formie deep fakes. Śledzenie proponowanych rozwiązań oraz aktualnej debaty może być istotnym elementem budowania kompetencji legislacyjnych oraz technicznych także w Polsce, zwłaszcza wobec zgłaszanych w 2025 r. koncepcji penalizacji nowych czynów związanych z upowszechnianiem CSAM w formie deep fakes, które na ten moment mogą nastręczać trudności z kwalifikacją na gruncie polskiego Kodeksu karnego (np. instrukcje tworzenia zakazanych treści z wykorzystaniem modeli AI). Zgłaszanie propozycji legislacyjnych powinno uwzględniać bogatą aktywność Dyżurnet.pl oraz przegląd globalnych trendów,

np. z uwzględnieniem corocznego raportu INHOPE „Global CSAM Legislative Overview 2024.", w którym od 2024 r. uwzględniana jest kategoria CSAM wygenerowanego przez AI[49]. W tym kontekście zalecane jest stałe wspieranie Dyżurnet.pl i NASK w bieżącej pracy i zapewnianie najlepszych możliwych warunków finansowych i technicznych. Wymiar sądownictwa powinien śledzić precedensowe orzeczenia sądów w innych krajach, wykładnię przepisów oraz uzasadnienia towarzyszące wyrokom. Niezależnie od systemu prawnego i obowiązujących przepisów prawnych znaczna część orzeczeń ma charakter uniwersalny, wskazując na pożądane z prawnego i społecznego punktu widzenia linie orzecznicze dostosowane do dynamicznego rozwoju nowoczesnych technologii, w tym generatywnej AI.

49 INHOPE (2024). Global CSAM Legislative Overview 2024. INHOPE: Amsterdam.