



ŚMIERĆ

PRZEZ TYSIĄC CIĘĆ

Raport o wzmacnianej technologicznie
przemocy wobec społeczności
LGBTI na platformie X

AMNESTY
INTERNATIONAL



Amnesty International to globalny ruch zrzeszający ponad 10 milionów osób działających na rzecz świata, w którym wszyscy mogą korzystać z praw człowieka.

Chcemy zapewnić każdemu możliwość realizacji praw zapisanych w Powszechnej Deklaracji Praw Człowieka oraz innych dokumentach międzynarodowych poświęconych prawom człowieka.

Jesteśmy niezależni od władz państwowych, ideologii politycznych, interesów gospodarczych lub religii, a fundusze na swoją działalność pozyskujemy w większości ze składek członkowskich i dotacji osób indywidualnych.

© Amnesty International 2019 O ile nie wskazano inaczej, zawartość niniejszego dokumentu objęta jest licencją Creative Commons (uznanie autorstwa, użytek niekomercyjny, bez utworów zależnych, międzynarodowe 4.0) <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> Więcej informacji o zgodzie na wykorzystanie materiałów dostępne na stronie: www.amnesty.org Materiały, których prawa autorskie przypisane są stronom innym niż Amnesty International, nie są objęte licencją Creative Commons.

Wydanie pierwsze w 2019 r.

Amnesty International Ltd
Peter Benenson House, 1 Easton Street
Londyn WC1X 0DW, W.B.

Index: EUR 37/0418/2019

Język oryginału: Angielski

amnesty.org

Zdjęcie na okładce: WARSZAWA, POLSKA - 22 lipca 2017: Protestujący gromadzą się przed siedzibą Sądu Najwyższego aby wezwać Prezydenta do zawetowania ustaw ograniczających niezależność sądownictwa. (c) Madeleine Lenz/Pacific Press/LightRocket via Getty Images

**AMNESTY
INTERNATIONAL**



SPIS TREŚCI

1. STRESZCZENIE	7
2. METODOLOGIA.....	12
2.1 EKSPERYMENT ILOŚCIOWY.....	12
2.2 KATEGORYZOWANIE TREŚCI	13
2.3 WYZWANIA METODOLOGICZNE	14
3. TŁO	16
3.1 CEL NAGONKI: SPOŁECZNOŚĆ LGBTI W POLSCE.....	16
3.2 „STREFY WOLNE OD LGBT”	18
3.3. BRAK OCHRONY PRAWNEJ.....	19
3.4 ZMIANA SYTUACJI POLITYCZNEJ.....	20
3.5 STATUS PLATFORMY X NA POLSKIEJ SCENIE POLITYCZNO-MEDIALNEJ	22
4. UWARUNKOWANIA PRAWNE	23
4.1. BIZNES I STANDARDY PRAW CZŁOWIEKA	23
4.2 PRAWA OSÓB LGBTI W KONTEKŚCIE BIZNESU.....	24
4.3 NALEŻYTA STARANNOŚĆ W ZAKRESIE PRZESTRZEGANIA PRAW CZŁOWIEKA A FIRMY Z BRANŻY TECHNOLOGICZNEJ	25
4.4 ODPOWIEDZIALNOŚĆ PRZEDSIĘBIORSTW W ZAKRESIE ŚRODKÓW NAPRAWCZYCH	27
4.5 ZOBOWIĄZANIA WYNIKAJĄCE Z AKTU O USŁUGACH CYFROWYCH (DSA)	27
4.6 PRAWA DO ŻYCIA WOLNEGO OD PRZEMOCY ZE WZGLĘDU NA PŁEĆ	28
4.7 ZAKAZ NAWOŁYWANIA DO NIENAWIŚCI W MIĘDZYNARODOWYCH PRAWACH CZŁOWIEKA	29

5. ROLA SERWISU X W PROPAGOWANIU PRZEMOCY ZE WZGLĘDU NA PŁEĆ WSPOMAGANEJ TECHNOLOGICZNIE W POLSCE.	31
5.1 PRZEMOC ZE WZGLĘDU NA PŁEĆ WSPOMAGANA TECHNOLOGICZNIE W SERWISIE X. ...	31
5.2 ROLA OSÓB PUBLICZNYCH I POLITYKÓW_POLITYCZEK W ROZPOWSZECZNIANIU TREŚCI ANTY-LGBTI W POLSCE.	33
5.3 ANALIZA: JAK PRZEMOC ZE WZGLĘDU NA PŁEĆ WSPOMAGANA TECHNOLOGICZNIE W POLSCE PRZEKŁADA SIĘ NA REALNE KRZYWDY	34
5.4 ZWOLNIENIA PERSONELU ODPOWIEDZIALNEGO ZA MODERACJĘ TREŚCI.	36
5.5 ZMIANY PODEJŚCIA SERWISU X DO TREŚCI NIENAWISTNYCH.	40
5.6 OPÓR PRZED DOSTOSOWANIEM SIĘ DO PRZEPISÓW I STANDARDÓW UE.	42
5.7 PROBLEMY SYSTEMOWE	43
 6. MIEĆ SIĘ NA BACZNOŚCI: WPŁYW PRZEMOCY ZE WZGLĘDU NA PŁEĆ WSPOMAGANEJ TECHNOLOGICZNIE NA OSOBY LGBTI	48
6.1 HISTORIA ALEKSANDRY HERZYK.	48
6.2 HISTORIA ALEGO	49
6.3 HISTORIA MAGDY DROPEK	50
6.4 HISTORIA MAI HEBAN.	51
6.5 HISTORIA NATHANA BRYZY.	51
 7. BIZNES NIENAWIŚCI: JAK MODEL BIZNESOWY X ZWIĘKSZA RYZYKO NARUSZANIA PRAW CZŁOWIEKA I KRZYWDZI LUDZI.	53
7.1 MODEL BIZNESOWY OPARTY NA INWIGILACJI	53
7.2 ALGORYTMY OPARTE NA ZAANGAŻOWANIU I ARCHITEKTURA SYSTEMU REKOMENDACJI X56	
7.3 RYZYKO ZWIĄZANE Z ALGORYTMAMI BAZUJĄCYMI NA ZAANGAŻOWANIU	60
7.4 KABINY POGŁOSOWE	62
7.5 PRIORYTETOWE TRAKTOWANIE IDEI „CYFROWEJ AGORY” KOSZTEM ŚRODKÓW ŁAGODZĄCYCH	65
7.6 BRAK KONTAKTÓW X ZE SPOŁECZEŃSTWEM OBYWATELSKIM	69
7.7 NIEWYSTARCZAJĄCE ŁAGODZENIE RYZYKA SYSTEMOWEGO	69
7.8 ŚWIADOMOŚĆ X W KWESTII ZAGROŻEŃ SYSTEMOWYCH	71

7.9 OCENA PRZYCZYNIENIA SIĘ X DO AKTÓW TFGBV PRZECIWKO SPOŁECZNOŚCI LGBTI W POLSCE	72
8. ZADOŚĆUCZYNIENIE I UNIKANIE SZKÓD W PRZYSZŁOŚCI	74
8.1 ODPOWIEDZIALNOŚĆ X ZA DZIAŁANIA NAPRAWCZE	74
8.2 WYPEŁNIANIE PRZEZ X OBOWIĄZKÓW WYNIKAJĄCYCH Z DSA	75
8.3 KARY PRZEWIDZIANE W DSA	76
9. WNIOSKI I ZALECENIA	77
9.1 WNIOSKI	77
9.2 ZALECENIA	78
9.2.1 ZALECENIA DLA FIRMY X.	78
9.2.1 ZALECENIA DLA KOMISJI EUROPEJSKIEJ	80
9.2.2 ZALECENIA DLA RZĄDU RZECZPOSPOLITEJ POLSKIEJ	80

WYKAZ SKRÓTÓW

Skrót	Opis
AI	sztuczna inteligencja
ATI	Algorithmic Transparency Institute
CEDAW	Komitety ONZ ds. Likwidacji Dyskryminacji Kobiet
DSA	Akt o usługach cyfrowych
DSC	koordynator ds. usług cyfrowych
Wytyczne OECD dotyczące należytej staranności	OECD Due Diligence Guidance for Responsible Business Conduct
ETPCz	Europejski Trybunał Praw Człowieka
EU	Unia Europejska
FTC	Federalna Komisja Handlu
GBV	przemoc ze względu na płeć
GNI	Global Network Initiative
ICCPR	Międzynarodowy Pakt Praw Obywatelskich i Politycznych
ISD	Institute for Strategic Dialogue
LLM	duży model językowy
Wytyczne OECD	wytyczne OECD dla przedsiębiorstw wielonarodowych dotyczące odpowiedzialnego prowadzenia działalności biznesowej
OHCHR	Wysoki Komisarz Narodów Zjednoczonych ds. Praw Człowieka
SOGIESC	orientacja seksualna, tożsamość płciowa, ekspresja płciowa i cechy płciowe
TFGBV	przemoc ze względu na płeć wspomagana technologicznie
Wytyczne dotyczące biznesu i praw człowieka	Wytyczne dotyczące biznesu i praw człowieka: wdrażanie dokumentu ramowego ONZ Chronić, Szanować i Naprawiać
VLOP	bardzo duża platforma internetowa

1. STRESZCZENIE

Ostrzeżenie o niepokojących treściach

W raporcie poruszane są trudne tematy, takie jak przemoc ze względu na płeć wspomagana technologicznie (TFGBV). Przedstawione zostaną przykłady drastycznych treści, między innymi nawoływania do przemocy i dyskryminacji, co może wywoływać niepokój u niektórych osób czytających raport.

„Twitter to właściwie potok deadnamingu, misgenderowania, wyzwisk i grożenia śmiercią”, powiedziała Maja Heban w wywiadzie udzielonym Amnesty International. Opisywała swoje doświadczenia jako transpłciowej kobiety korzystającej z platformy X (dawniej Twitter). Wszystkie inne osoby ze społeczności LGBTI, z którymi Amnesty International przeprowadzała wywiady na użytek tego raportu, powtarzały, że na platformie X treści wrogie wobec społeczności LGBTI są wszechobecne.

Polska społeczność LGBTI od kilkadziesiąt lat zмага się z systemową dyskryminacją. Nasiliła się ona w latach 2015–2023, czyli w okresie rządów Prawa i Sprawiedliwości, kiedy władze podejmowały działania mające na celu ograniczanie rozwoju społeczeństwa obywatelskiego, występowały przeciwko praworządności, atakowały prawa kobiet i osób LGBTI, a także tworzyły coraz mniej przyjazne środowisko dla osób LGBTI oraz ich sojuszników.

Wrogi, stygmatyzujący sposób mówienia o osobach LGBTI stał się powszechny. Tego rodzaju retoryką posługiwali się również politycy zajmujący najważniejsze stanowiska w państwie. W 2022 roku Amnesty International zgromadziła jednoznaczne dowody na to, że retoryka ta prowadziła do przemocy i do większej liczby ataków na osoby LGBTI, na przykład uczestniczące w pokojowych zgromadzeniach, marszach równości lub protestach.

Głośnym tego przykładem był Marsz Równości w Białymstoku w 2019 roku, którego uczestnicy zostali obrzuceni butelkami, kostką brukową i petardami przez kontrademonstrantów skandujących obraźliwe hasła. Kilka miesięcy później, podczas Marszu Równości w Lublinie, policja aresztowała kilkadziesiąt kontrademonstrantów, którzy zamierzali zaatakować uczestników pokojowego zgromadzenia. Ujawniono później, że dwoje kontrademonstrantów miało przy sobie materiały wybuchowe domowej roboty.

W 2020 roku wrogość wobec osób LGBTI w Polsce była tak skrajna, że około jednej trzeciej regionów kraju przyjęło symboliczne uchwały przeciwko „ideologii LGBT”.

Tymczasem platformę X zalewały wpisy podżegające do dyskryminacji osób LGBTI, podsycające nienawiść do tej społeczności lub wprost wzywające do agresji wobec niej. Spełniało to znamiona przemocy ze względu na płeć wspomaganej technologicznie (TFGBV), stanowiącej wielowymiarowe zagrożenie dla praw człowieka. Treści, o których tu mowa, często pojawiały się na kontach polityków i polityczek korzystających z platformy X. Wpisy publikowane przez wielu z nich wzywały do nienawiści lub dehumanizowały osoby LGBTI, sugerując, że ich tożsamość to „ideologia” i że stanowią one zagrożenie dla bezpieczeństwa dzieci. Z powodu upowszechnienia się takich wpisów na platformie X podsycanie nienawiści do osób LGBTI było normalizowane i zyskiwało coraz większe przyzwolenie społeczne.

Wpisy noszące znamiona TFGBV zalewały platformę X po części ze względu na to, że nie podejmowała ona adekwatnych działań w celu moderowania treści. Sytuacja pogorszyła się jeszcze bardziej po tym, jak w październiku 2022 roku Elon Musk przejął serwis i przeprowadził masowe zwolnienia. Zaledwie tydzień później użytkownicy i użytkowniczki propagujący fundamentalistyczne narracje zaczęli sprawdzać, jak daleko mogą się posunąć w mowie nienawiści wobec osób LGBTI. Były zawodnik mieszanych sztuk walki Jake

Shields (obserwowany przez 34 tysiące użytkowników_użytkowniczek X) opublikował zdjęcie drag queen z podpisem: „Tak wygląda groomer”. Następnie dodał: „Dwa miesiące temu zawiesili mnie za tego samego tweeta. Zobaczmy, czy teraz na Twitterze naprawdę panuje wolność”.

Krótko po przejściu X Elon Musk rozwiązał Radę ds. Zaufania i Bezpieczeństwa, czyli ciało doradcze, w którego skład wchodziło 100 organizacji społecznych, między innymi organizacji działających na rzecz praw człowieka, utworzone po to, by przeciwdziałać treściom mogącym prowadzić do przemocy wobec dzieci, samobójstw, samookaleczeń i upowszechniania się mowy nienawiści. Szacuje się, że Elon Musk zwolnił ponadto 80% personelu odpowiedzialnego za kwestie zaufania i bezpieczeństwa. Pod koniec 2022 roku media doniosły, że Musk chce, aby moderowanie treści odbywało się w X głównie z wykorzystaniem narzędzi cyfrowych, choć, jak powszechnie wiadomo, prowadzi to do licznych błędów i nadużyć. W 2023 roku platforma wprowadziła funkcję „Uwagi społeczności” (Community Notes), cedując tym samym część zadań w zakresie moderowania treści na losowo wybranych, zweryfikowanych użytkowników_użytkowniczek.

Pod kierownictwem Elona Muska zmieniła się też polityka X w zakresie szkodliwych treści, w tym treści spełniających znamiona TFGBV. Na przykład w kwietniu 2023 roku zdecydowano, że „umyślne misgenderowanie lub deadnaming osób transpłciowych” nie będą już naruszały warunków korzystania z platformy. Decyzja ta została cofnięta w 2024 roku.

Przed przejściem X Musk zapowiadał, że zamierza znieść większość ograniczeń dotyczących treści umieszczanych na platformie i sugerował, że użytkownicy_użytkowniczek powinni mieć pełną swobodę wypowiedzenia się, o ile nie naruszają przepisów obowiązujących w krajach, w których działa dostawca usługi X.

Wydaje się, że spełnił swoje obietnice. Do 30 października 2023 roku *Zasady użytkowania X* obejmowały zapis: „Prowadzimy **politykę zero tolerancji** [podkreślenie: Amnesty International], jeśli chodzi o mowę nienawiści, żeby zapewnić bezpieczeństwo naszym użytkownikom_użytkowniczkom i zapobiec normalizowaniu przemocy”. Po zmianach zapis brzmi: „**Możemy usuwać niektóre treści lub ograniczać ich widoczność** [podkreślenie: Amnesty International], żeby zapewnić bezpieczeństwo naszym użytkownikom_użytkowniczkom i zapobiec normalizowaniu przemocy”.

Przedstawiciele_przedstawicielki społeczności LGBTI w Polsce mówili Amnesty International, że bycie widocznymi na platformie X naraża ich na zalew nienawistnych treści ze względu na ich prawdziwą lub postrzeganą płeć, orientację seksualną oraz tożsamość i ekspresję płciową. Wiele osób, z którymi przeprowadzono wywiady, przyznaje, że nienawistna retoryka, z którą spotykają się online, ma negatywny wpływ na ich dobrostan psychiczny. Joanna Prochowicz, lesbijka mieszkająca w Lublinie, powiedziała Amnesty International: „Powinniśmy traktować media społecznościowe jak część naszego życia społecznego. To, co piszemy w internecie, może krzywdzić tak samo jak słowa wypowiedziane w rozmowie. [...] To szkodliwe, bolesne, czasami bardzo. Media społecznościowe nie tyle wpływają na nasze normalne życie, ile po prostu są jego częścią. I dlatego mają na nas wpływ”.

Aleksandra Herzyk, asekualna kobieta mieszkająca w Krakowie, powiedziała Amnesty International, że padła ofiarą ataków ze strony użytkowników_użytkowniczek X, bo pisała na platformie o swojej asekualności. Doświadczyła też napaści spełniającej znamiona TFGBV, kiedy napisała o zabiegu zmniejszania piersi, na który się zdecydowała – niektórzy użytkownicy błędnie uznali, że jest transpłciową kobietą. „[...] czytasz różne rzeczy na swój temat, rzeczy, które nie są prawdą, ale jakoś zostają ci w głowie. To jak śmierć przez tysiąc cięć” – komentuje Aleksandra

Po doświadczeniach z mową nienawiści w X Herzyk przestała korzystać z platformy. Usunęła konto na początku 2024 roku. W 2018 roku w raporcie *Toxic Twitter* [Toksyczny Twitter] Amnesty International wykazała, że X (wówczas Twitter) nie chroni prawa kobiet do bezpieczeństwa w sieci, ponieważ nie podejmuje stosownych kroków w kierunku przeciwdziałania przemocy online, dotyczącej zwłaszcza kobiety należące do mniejszości etnicznych, rasowych lub religijnych, kobiety z niepełnosprawnościami lub identyfikujące się jako lesbijki, osoby biseksualne, niebinarne lub transpłciowe. W 2020 roku Amnesty ustaliła, że od 2018 roku X poczynił pewne postępy w przeciwdziałaniu TFGBV, lecz mimo to nadal w niedostatecznym stopniu chroni prawa człowieka osób korzystających z platformy.

Wygłąda na to, że od 2020 roku niewiele się poprawiło – przynajmniej na gruncie polskim. W 2024 roku Stowarzyszenie „Nigdy Więcej” opublikowało raport, w którym udokumentowało 343 przykłady mowy nienawiści zgłoszone administratorom platformy między sierpniem 2023 roku a sierpniem 2024 roku. Stowarzyszenie „Nigdy Więcej” ma status zaufanego zgłaszającego w ramach projektu wspieranego finansowo przez program Unii Europejskiej Obywatele, Równość, Prawa i Wartości. W zdecydowanej większości przypadków X zignorował zgłoszenia lub odmówił usunięcia wpisów, także tych podlegających do nienawiści i dyskryminacji marginalizowanych mniejszości, w tym społeczności LGBTI. Wiele postów zgłoszonych przez Stowarzyszenie „Nigdy Więcej” – między innymi wpisy, w których osoby LGBTI są

przedstawiane jako dewianci i w których pojawiają się wyzwiska, obraźliwe sformułowania, nawoływanie do nienawiści, a nawet eksterminacji – nadal można znaleźć na platformie X.

W niniejszym raporcie Amnesty International wykazuje, że X nie chroni polskiej społeczności LGBTI przed przemocą online, w znacznym stopniu dlatego że treści zamieszczane na platformie nie są odpowiednio moderowane, a platforma nie przykładą dostatecznie dużej wagi do kwestii związanych z prawami człowieka. Tym samym X przyczynia się do naruszeń praw człowieka dotyczących społeczności LGBTI w Polsce. Przeznaczanie niedostatecznych zasobów na moderowanie treści stanowiło problem, zanim jeszcze w 2022 roku Elon Musk przejął firmę. Ponadto X nie prowadzi odpowiednio ścisłej współpracy z organizacjami LGBTI w Polsce, co pozwoliłoby przeciwdziałać zagrożeniom dla tej społeczności wynikającym z treści publikowanych na platformie. Ta niewydolność – w połączeniu z niemożliwą do usprawiedliwienia decyzją o rezygnacji z narzędzi chroniących użytkowników_użytkowniczki przed mową nienawiści i przemocą oraz z „absolutyzmem wolności słowa” deklarowanym przez Elona Muska – sprawia, że platforma X jest pełna treści spełniających znamiona TFGBV, w tym nawoływania do nienawiści, przemocy i dyskryminacji dotyczących osoby LGBTI.

Amnesty International we współpracy z National Conference on Citizenship’s Algorithmic Transparency Institute (ATI) przeprowadziła między 1 a 31 marca 2025 roku badanie ilościowe z wykorzystaniem 32 kont marionetkowych na platformie X, w ramach którego zgromadzono 163 048 wpisów¹. Badanie wykazało, że treści wymierzone w społeczność LGBTI są na X bardzo rozpowszechnione. Z analizy próbek 1387 wpisów wynika, że treści homofobiczne i transfobiczne szczególnie często spotyka się w feedach użytkowników_użytkowniczek obserwujących osoby ze świata polityki sprzeciwiające się prawom osób LGBTI. Amnesty International wykazała, że prawie 4% treści zebranych dzięki kontom marionetkowym, które posłużyły do obserwowania osób ze świata polityki sprzeciwiających się prawom osób LGBTI, spełniało znamiona homofobii lub transfobii. Ponadto ponad 25% wszystkich wpisów dotyczących tematyki LGBTI zebranych dzięki tym kontom nosiło znamiona homofobii lub transfobii. Amnesty International ustaliła też, że znaczny odsetek treści dotyczących tematyki LGBTI (wpisów lub odpowiedzi na wpisy) miał charakter homofobiczny lub transfobiczny. Treści takie często pojawiały się również w feedach kont marionetkowych służących do obserwowania polityków_polityczek wspierających prawa osób LGBTI.

W latach 2015–2023 w Polsce rządziło Prawo i Sprawiedliwość, jawnie występujące przeciwko społeczności LGBTI. Co prawda po wyborach w 2023 roku władzę przejęły inne partie, ale wieloletnia nagonka na społeczność LGBTI doprowadziła do sytuacji, którą aktywiści_aktywistki określają mianem „odgórnej polaryzacji”. Odzwierciedleniem tego są treści wymierzone w społeczność LGBTI publikowane na platformie X. Szczególny niepokój budzi ich powszechność, gdyż model biznesowy X opiera się na rekomendowaniu treści, które przyciągają uwagę użytkowników_użytkowniczek, bez względu na to, czy mogą powodować szkody.

W niniejszym raporcie Amnesty International przedstawia pierwszą całościową analizę modelu biznesowego X z perspektywy praw człowieka. Wykazujemy, że w przypadku tej platformy mamy do czynienia z modelem opartym na inwigilacji i gromadzeniu danych o użytkownikach_użytkowniczkach, tak samo jak w przypadku firm Meta, Google czy TikTok. Gromadzenie danych stanowi najważniejszy element modelu biznesowego X, bo po pierwsze pozwala lepiej przewidywać, jakie treści przyciągną uwagę użytkowników_użytkowniczek, a po drugie z perspektywy potencjalnych reklamodawców wartość danych przekłada się na wartość samej firmy. To kwestia kluczowa, gdyż X może działać wyłącznie dzięki publikowaniu targetowanych reklam.

Od 2013 roku targetowane reklamy odpowiadają za prawie 100% przychodów firmy. Aby zoptymalizować gromadzenie danych o użytkownikach_użytkowniczkach, algorytmy X czynią przyciąganie uwagi absolutnym priorytetem. W rezultacie największą widoczność zyskują treści, które z największym prawdopodobieństwem wywołają jakiegoś rodzaju reakcję (w przypadku X reakcją może być komentarz, polubienie lub retweet). X oferuje również subskrypcję premium: użytkownicy_użytkowniczki płacą za dodatkowe funkcjonalności, na przykład za możliwość publikowania dłuższych wpisów i podkreślanie algorytmu obejmujące „priorytet odpowiedzi” – oznacza to, że odpowiedzi użytkowników_użytkowniczek premium są bardziej widoczne.

Amnesty International już wcześniej wykazała, że modele biznesowe oparte na inwigilacji prowadzą do rozpowszechniania szkodliwych treści, gdyż stosującym je platformom cyfrowym zależy głównie na przyciąganiu uwagi i utrzymywaniu zainteresowania użytkowników_użytkowniczek oraz na gromadzeniu danych. W połączeniu z kiepską moderacją treści model biznesowy tego typu realizowany przez X naraża polską społeczność LGBTI na wyjątkowo duże zagrożenie treściami wypełniającymi znamiona TFGBV.

¹ Konta marionetkowe to konta fikcyjnych użytkowników_użytkowniczek tworzone na użytek badań. Można je wykorzystywać na różne sposoby. W tym przypadku Amnesty International i ATI posłużyły się kontami marionetkowymi, żeby przeanalizować rozpowszechnienie i nagłaśnianie treści wymierzonych przeciwko społeczności LGBTI na platformie X w Polsce.

Za typowy przykład treści wymierzonych w społeczność LGBTI i rozpowszechnianych na platformie X może posłużyć klip z lipca 2023 roku opublikowany przez Konfederację. Zawiera on fragment wypowiedzi jej ówczesnego posła Grzegorza Brauna, który przemawiając w sejmie na temat społeczności LGBTI, mówi: „Nie będą dewianci uczyć naszych dzieci tolerancji. Precz z propagandą sodomii w polskich szkołach”. W maju 2025 roku klip nadal był dostępny w serwisie X. Miał ponad 99 tysięcy wyświetleń.

Osoby LGBTI mówiły Amnesty International, że regularnie widzą na platformie X wpisy, które je dehumanizują, a nawet wzywają do ich eksterminacji. „Osoby LGBTQ mają skończyć w komorach gazowych, mówi się o nas, jakbyśmy byli śmieciami, brudem, który trzeba oczyścić”. Inna rozmówczyni Amnesty International przytaczała wpisy o tym, że „[osoby LGBTI] nie są normalne, są przeciwko polskim rodzinom, chcą je zniszczyć; to nie ludzie, to ideologia”.

X przeznacza zdecydowanie niewystarczające zasoby na moderowanie treści, również w Polsce, co stanowi istotną przyczynę tego, że platforma nie usuwa treści spełniających znamiona TFGBV wymierzonych w polską społeczność LGBTI. Według raportów publikowanych przez firmę X zatrudnia tylko dwie polskojęzyczne osoby odpowiedzialne za moderowanie treści, przy czym w przypadku jednej z nich polski nie jest językiem ojczystym. Mówimy tu o dwóch osobach na 37,45 miliona mieszkańców kraju i 5,33 miliona użytkowników_użytkowniczek X, co dobrze pokazuje, że firma nie inwestuje w moderowanie treści. Innym przykładem jest wprowadzenie „Uwag społeczności”, czyli *de facto* scedowanie moderowania treści na użytkowników_użytkowniczki. Niewystarczające zasoby oraz polityka i konkretne działania firmy przyczyniają się do tego, że na platformie X pełno jest treści wymierzonych w społeczność LGBTI.

Każda firma ma obowiązek respektować prawa człowieka, bez względu na to, w której części świata prowadzi działalność. Aby wypełniać ten obowiązek, firmy muszą być proaktywne i nieustannie monitorować wpływ swoich usług na realizację praw człowieka, identyfikować potencjalne naruszenia i negatywne skutki swojej działalności, zapobiegać im, a jeśli do nich doszło, stosować odpowiednie środki naprawcze. W przypadku firm z branży technologicznej takich jak X konieczne jest też reagowanie w sytuacjach, w których model biznesowy, działalność, decyzje projektowe i praktyki w zakresie moderowania treści powodują lub zwiększają ryzyko naruszeń praw człowieka.

Zgodnie z Aktem o usługach cyfrowych (DSA) obowiązującym w Unii Europejskiej tak zwane bardzo duże platformy internetowe (VLOP), do których zalicza się X, są zobowiązane szacować ryzyko systemowe i dążyć do jego zmniejszania, a także przygotowywać coroczne sprawozdania w tym zakresie. Najnowsze dostępne sprawozdanie przygotowane przez X pochodzi z 2024 roku. Firma przyznała, że treści publikowane przez użytkowników_użytkowniczki X mogą spełniać znamiona mowy nienawiści lub przemocy wobec pewnych grup i jednostek, co z kolei może powodować lęk i uczucie zastraszenia, a także prowadzić do autocenzury. X wskazał środki, które mają zredukować to zjawisko: należą do nich między innymi ograniczenie widoczności pewnych treści, przejrzystość zasad i procedur, wprowadzenie procedur dotyczących kontroli jakości oraz ewaluacja zasad. W sprawozdaniu nie wspomniano jednak o konkretnym ryzyku, na które narażona jest społeczność LGBTI. Niezależny audyt przeprowadzony na mocy DSA, obejmujący okres do 23 sierpnia 2024 roku, wykazał, że proces szacowania ryzyka systemowego w X jest nie dość rygorystyczny, a środki wskazane przez platformę nie wystarczają, żeby to ryzyko ograniczyć. Podkreślono między innymi, że środki te nie obejmują żadnych modyfikacji algorytmu.

W niniejszym raporcie wykazano, że X nie dokłada odpowiednich starań, by monitorować swój wpływ na przestrzeganie praw człowieka w Polsce, mimo obowiązku szacowania ryzyka, który to obowiązek został ustanowiony na mocy DSA. Platforma nie poczyniła kroków zmierzających do zapobiegnięcia szkodom stanowiącym skutek jej działalności. Jest niepodważalnym faktem, że X przyczynia się do upowszechniania treści spełniających znamiona TFGBV i do naruszeń praw człowieka dotyczących polską społeczność LGBTI.

22 sierpnia 2024 roku Amnesty International skontaktowała się z X. Zadała pytania dotyczące działań dostawcy usług w Polsce w latach 2019–2024. X nie udzielił odpowiedzi.

W niniejszym raporcie Amnesty International wykazuje, że platforma X nie wypełnia swoich podstawowych obowiązków w zakresie ochrony praw człowieka wynikających z *Wytycznych dotyczących biznesu i praw człowieka* przyjętych przez Organizację Narodów Zjednoczonych, a ponadto nie spełnia wymogów stawianych przez DSA. Jedno i drugie sprawia, że polska społeczność LGBTI ponosi znaczące szkody. Działania zapobiegawcze i naprawcze prowadzone przez X są nieadekwatne. Platforma lekceważy treści promujące przemoc i nienawiść, a jej model biznesowy zdecydowanie zwiększa systemowe ryzyko naruszeń praw człowieka. Z wymienionych tu powodów jest szczególnie prawdopodobne, że w wyniku funkcjonowania X polska społeczność LGBTI będzie ponosiła dalsze szkody. Konieczne są pilne, szeroko zakrojone zmiany, które zagwarantują, że X przestanie przyczyniać się do tych szkód. Do niezbędnych i gruntownych zmian zaliczają się przeznaczenie odpowiednich zasobów na moderowanie treści oraz odejście od modelu biznesowego opartego na inwigilacji.

Z bieżących obserwacji funkcjonowania serwisu X w Polsce wynika, że nadal nie poczynił on kroków, żeby zminimalizować systemowe ryzyko naruszeń praw człowieka. DSA zawiera konkretne rozwiązania, które zwiększają rozliczalność i mogą zmusić platformy do wdrożenia środków naprawczych, powinien więc zostać w pełni zaimplementowany.

Niestety, jak dotąd polski rząd nie wprowadził wszystkich niezbędnych zmian ustawowych: kompetencje powołanego tymczasowo koordynatora ds. usług cyfrowych ograniczone są do przepisów organizacyjno-technicznych aktu o usługach cyfrowych, nie ustanowiono też kar za naruszanie DSA. Powołanie koordynatora ds. usług cyfrowych i przyznanie mu odpowiednich prerogatyw, a także zasobów – zwłaszcza funduszy i dostępu do wiedzy eksperckiej – jest kwestią o pierwszorzędym znaczeniu. Dopóki się to nie stanie, polscy użytkownicy i polskie użytkowniczki X nie będą mogli w pełni korzystać z praw gwarantowanych im przez DSA. W maju 2025 roku Komisja Europejska skierowała do Trybunału Sprawiedliwości Unii Europejskiej sprawę przeciwko Polsce, a także Czechom, Hiszpanii, Cyprowi i Portugalii, w związku z niezapewnieniem skutecznego wdrożenia aktu o usługach cyfrowych.

Być może Komisja Europejska powinna wszcząć śledztwo w sprawie X dotyczące tego, w jaki sposób platforma stara się minimalizować ryzyko systemowe wynikające zarówno z jej modelu biznesowego, jak i z praktyk w zakresie moderowania treści. Kwestia ta ma szczególne znaczenie ze względu na utrzymujący się negatywny wpływ treści spełniających znamiona TFGVB publikowanych na platformie X na polską społeczność LGBT – w tym na prawo jednostek do wolności słowa i wyrażania siebie, a także prawo do ochrony przed dyskryminacją.

Unia Europejska ma odpowiednie narzędzia, żeby wypełniać swoje zobowiązania w zakresie ochrony praw człowieka – w tym prawa do życia wolnego od przemocy ze względu na płeć. Pora, żeby zaczęła wykorzystywać te narzędzia w pełni i konsekwentnie.

2. METODOLOGIA

Niniejszy raport bazuje na badaniach przeprowadzonych przez Amnesty International między marcem 2024 roku a kwietniem 2025 roku, obejmujących metody partycypacyjne, ilościowe (eksperyment badawczy opisany poniżej) oraz wywiady pogłębione. Amnesty International przeanalizowała, jakie konsekwencje dla przestrzegania prawa człowieka w Polsce ma działalność biznesowa platformy X, wykorzystując dokumenty dostępne na jej stronie internetowej oraz sprawozdania dotyczące przejrzystości przygotowywane przez X na mocy DSA. Raport został uzupełniony wywiadami eksperckimi, wynikami badań ilościowych na platformie X oraz analizą dokumentów. Amnesty International przeprowadziła ponadto szeroko zakrojony przegląd dostępnych badań, raportów organizacji pozarządowych oraz informacji pochodzących z mediów krajowych i zagranicznych.

Raport stanowi kontynuację wcześniejszych badań prowadzonych przez Amnesty International, z których wynika, że społeczność LGBTI w Polsce doświadcza ustawicznej dyskryminacji, a wroga, stygmatyzująca retoryka – stosowana między innymi przez wysokich przedstawicieli władz państwowych – ma coraz bardziej szkodliwe konsekwencje, obejmujące przemoc i dyskryminację ze względu na prawdziwą lub postrzeganą orientację seksualną oraz tożsamość i ekspresję płciową².

Amnesty International przeprowadziła wywiady – twarzą w twarz lub zdalne – z 11 osobami dotkniętymi przemocą online. W raporcie przedstawiono ich historie, stanowią one bowiem ilustrację krzywd, które dotyczą społeczności LGBTI w Polsce, również z powodu działań X. Ponadto przeprowadzono pięć wywiadów eksperckich, w tym ze specjalistami w zakresie praw cyfrowych i praw osób LGBTI w Polsce. Zorganizowano również warsztat partycypacyjny z udziałem 16 osób z polskiej społeczności LGBTI.

Wszystkie wywiady prowadzono głównie po angielsku. Jeśli było to konieczne, zapewniano tłumaczenie niektórych fragmentów na polski. Informacje zebrane na podstawie wywiadów zostały dodatkowo zweryfikowane dzięki danym uzyskanym od organizacji zajmujących się prawami cyfrowymi i organizacji LGBTI, a także dzięki badaniom ilościowym na platformie X wykonanym przez Amnesty International. Wszystkie osoby, z którymi przeprowadzano wywiady, udzieliły wcześniej stosowanych zgód. Amnesty International nie oferowała żadnej gratyfikacji w zamian za udział w wywiadzie. Ze względów bezpieczeństwa niektóre osoby zażyczyły sobie anonimowości, inne pragnęły wystąpić pod własnymi nazwiskami. W przypadku tych pierwszych Amnesty International użyła pseudonimów i pominęła w raporcie wszelkie dane umożliwiające ich identyfikację.

2.1 EKSPERYMENT ILOŚCIOWY

Wywiady indywidualne i badania partycypacyjne są niezbędne, żeby zrozumieć i wydobyć na pierwszy plan doświadczenia ludzi. Stanowią też główne narzędzie aktywizmu i budowania świadomości. Nieprzejrzystość mechanizmów wykorzystywanych przez X i rozpowszechnianie pewnego rodzaju treści na platformie wymagały jednak zastosowania w tym projekcie metod ilościowych. Uzupełnieniem wywiadów indywidualnych i badań partycypacyjnych jest zatem komponent ilościowy, przygotowany we współpracy z ATI.

W idealnym scenariuszu badacze, którzy chcą analizować, w jaki sposób algorytmy X podsuwają użytkownikom_użytkowniczkom określone treści – w tym wszechobecne na platformie treści skierowane przeciwko społeczności LGBTI – gromadziliby dane pochodzące od prawdziwych użytkowników_użytkowniczek. Amnesty International rozważała tę metodę, ustalono jednak, że ze względu na ograniczenia

² Amnesty International, *Dotknięci przez nienawiść, zapomniani przez prawo. Brak spójnego systemu zwalczania przestępstw z nienawiści w Polsce* (Indeks: EUR 37/2147/2015), wrzesień 2015, https://www.amnesty.org.pl/wp-content/uploads/2015/09/hate_crimes_report_-_POLISH_-_WEB.pdf; Amnesty International, *„Byliśmy traktowani jak przestępcy”. Od atmosfery wrogości po nękanie osób broniących praw LGBTI* (Indeks: EUR 37/5882/2022), 20 lipca 2022, https://www.amnesty.org.pl/wp-content/uploads/2022/07/Bylismy-traktowani-jak-przestepcy-raport-aktywisci_stki-LGBTI-Amnesty-International-2022.pdf.

czasowe zgromadzenie odpowiedniej ilości danych mogłoby się okazać niemożliwe. Najlepszą alternatywą było stworzenie kont marionetkowych, odwzorowujących konta użytkowników_użytkowniczek X.

Eksperyment ilościowy obejmował: (1) badanie rozpowszechnienia treści skierowanych przeciwko społeczności LGBTI na platformie X w Polsce i przeszukiwanie platformy pod kątem takich treści; (2) badanie, w jaki sposób algorytmy X przyczyniają się do większej widoczności szkodliwych treści. W tym celu zastosowano konta marionetkowe, a konkretnie konta fikcyjnych użytkowników_użytkowniczek potencjalnie zainteresowanych treściami skierowanymi przeciwko społeczności LGBTI. Eksperyment przeprowadzono między 1 a 31 marca 2025 roku.

Badacze utworzyli 32 konta marionetkowe, wykorzystując cztery różne, ustalone z góry osoby, aby w ten sposób odtworzyć doświadczenia różnych użytkowników_użytkowniczek platformy X. Poszczególne osoby obserwowały polityków_polityczki z określonych partii politycznych. Stworzono je w taki sposób, aby reprezentowały cztery różne podejścia do społeczności LGBTI (i innych kwestii społecznych). Każde z 32 kont marionetkowych obserwowało 10 polityków_polityczek reprezentujących jedną z poniższych kategorii:

- ☐ politycy_polityczki z partii opowiadających się za ograniczaniem praw osób LGBTI (PiS, Konfederacja, Kukiz15);
- ☐ politycy_polityczki z partii o niejednoznacznym stanowisku w kwestii praw osób LGBTI (PSL);
- ☐ politycy_polityczki z partii, które popierają związki partnerskie, ale sprzeciwiają się małżeństwom jedнопłciowym (Polska 2050, KO);
- ☐ politycy_polityczki z partii, które w pełni popierają prawa osób LGBTI (Nowa Lewica, Razem).

Aby badanie w jak najmniejszym stopniu przyczyniało się do rozpowszechniania potencjalnie szkodliwych treści, zdecydowano, że żadne z kont marionetkowych nie będzie służyło do dawania polubień, podawania dalej wpisów, komentowania, wysyłania wiadomości ani wyszukiwania jakichkolwiek słów czy haseł.

Każde konto marionetkowe logowało się na platformie X raz dziennie przez 28 dni. W połowie badania (15. dnia) każde konto przydzielono losowo do jednej z dwóch grup: (1) konta, które posłużą do obserwowania użytkowników_użytkowniczek z listy zarekomendowanych przez X; (2) konta, które nie będą obserwować rekomendowanych użytkowników_użytkowniczek. Codzienny workflow dla każdego konta marionetkowego wyglądał następująco:

1. Zalogować się na platformie X.
2. Spisać wszystkich użytkowników i wszystkie użytkowniczki osi czasu „Może Ci się spodobać”.
3. Przeszkrolować pierwszych 200 wpisów wyświetlonych w feedzie w kolejności od najnowszych do najstarszych.
4. Przeszkrolować pierwszych 200 wpisów w osi czasu „Dla Ciebie”.
5. Wylogować się z platformy X.

Konta posłużyły też do zebrania wszystkich komentarzy pod 55 wpisami dotyczącymi osób LGBTI mającymi najwyższe wskaźniki zaangażowania.

2.2 KATEGORYZOWANIE TREŚCI

Badanie Amnesty International było poświęcone (1) wszelkiego rodzaju skumulowanym efektom zwiększania widoczności treści skierowanych przeciwko osobom LGBTI przez platformę X; (2) konkretnym przypadkom nieskutecznego moderowania treści. Badanie upowszechnienia i zwiększania widoczności treści na platformach takich jak X zawsze wiąże się ze szczególnymi wyzwaniami, wymaga bowiem kategoryzowania określonych wpisów i treści.

We współpracy z Amnesty International Polska badacze_badaczki wypracowali zasady kategoryzowania poszczególnych wpisów. Wykorzystano tu zalecenia eksperckie, ale przyjęta kategoryzacja z konieczności upraszcza rzeczywistość i analizowane treści, a także ich potencjał wyrządzania ludziom krzywdy. Zastosowana kategoryzacja miała zagwarantować przejrzysty, wielowymiarowy opis potencjalnie szkodliwych treści skierowanych przeciwko osobom LGBTI, publikowanych na platformie X.

Próba 1387 wpisów została skategoryzowana ręcznie przez badaczy_badaczki. Posłużono się następującymi kategoriami:

1. przemocowy,
2. homofobiczny,
3. transfobiczny,
4. wykorzystywanie wizerunku intymnego bez zgody,
5. doxing,
6. nękanie,
7. inny problem,
8. antyimigrancki,
9. antysemitki,
10. codzienny rasizm,
11. ogólnie anti-LGBTI,
12. ogólnie pro-LGBTI,
13. LGBTI – inny,
14. nieistotny.

Następnie wykorzystano duży model językowy, żeby skategoryzować odpowiedzi pod 55 wpisami dotyczącymi osób LGBTI mającymi najwyższe wskaźniki zaangażowania. Model językowy został wyszkolony na podstawie ręcznie skategoryzowanych przykładów. Następnie skategoryzował odpowiedzi w zależności od tego, czy były homofobiczne/transfobiczne, czy też nie³.

Dzięki 32 kontom marionetkowym w okresie badania zebrano łącznie 163 048 wpisów pochodzących z feedów. 141 563 z nich nie było reklamami. Wiele wpisów się powtarzało, więc łączna liczba unikatowych wpisów jest niższa: 33 879, z czego 31 951 nie było reklamami.

Ponadto zgromadzono łącznie 10 455 unikatowych wpisów niebędących reklamami zamieszczonymi w osi czasu „Dla Ciebie”. Treści publikowane przez Elona Muska na jego osobistym koncie znacząco przeważały tu nad treściami z innych kont. W okresie badania w osi czasu „Dla Ciebie” na kontach marionetkowych wpisy Elona Muska wyświetliły się 1411 razy. Drugie miejsce zajęło konto polskiego polityka Romana Giertycha. Jego wpisy wyświetliły się 494 razy.

2.3 WYZWANIA METODOLOGICZNE

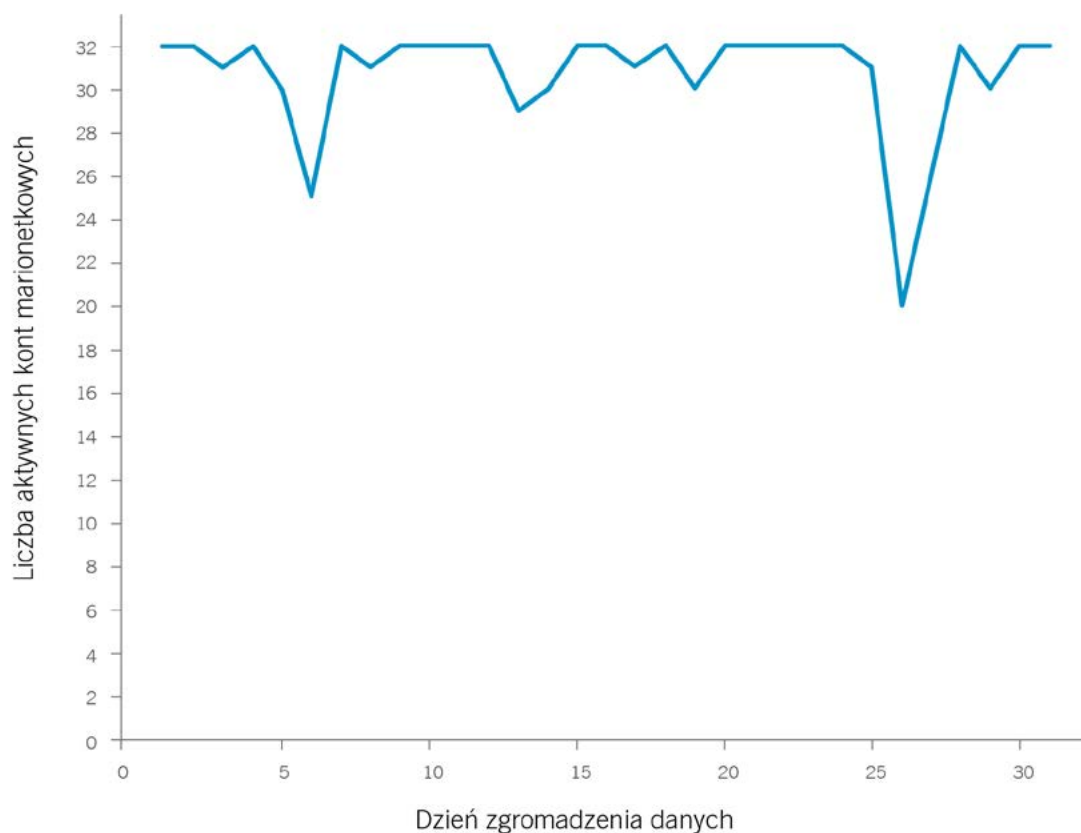
Badanie ilościowe obejmowało więcej wpisów niż poprzednie badania w tej tematyce prowadzone przez organizacje pozarządowe; mimo to 32 konta marionetkowe to za mało, żeby móc testować statystycznie jakiegokolwiek hipotezy, zdecydowano się więc jedynie na analizę opisową i porównanie ilości i częstości treści dotyczących osób LGBTI wyświetlanych każdej z czterech grup kont wyróżnionych na użytek badania, zwłaszcza treści szkodliwych. Rozwinięciem poprzednich badań było zastosowanie metody porównawczej, czyli grup kont obserwujących polityków i polityczki zajmujących różne stanowiska w kwestii praw osób LGBTI.

Jest wysoce prawdopodobne, że algorytm rekomendujący treści na platformie X uwzględnia takie czynniki jak data, godzina i lokalizacja użytkownika_użytkowniczki. Mogą one wpływać na to, jakie treści będą widoczne dla użytkowników_użytkowniczek. Aby zwiększyć użyteczność porównania, badacze_badaczki tworzyli konta marionetkowe w określone dni, o określonych godzinach, w tej samej lokalizacji na terenie Polski. Wyzwaniem okazały się problemy techniczne, na przykład awarie lub zamykanie kont marionetkowych. Badaczom_badaczkom udało się jednak w najwyższym możliwym stopniu zadbać o to, żeby dane w każdej grupie były gromadzone równocześnie.

³ Wykorzystano ChatGPT 4.1mini, 15 kwietnia 2025 roku. Użyto następującego promptu: „Jesteś binarnym klasyfikatorem, który decyduje, czy konkretna odpowiedź na tweeta jest homofobiczna/transfobiczna, czy też nie. Dla kontekstu dostaniesz oryginalny tweet i odpowiedzi, które zaklasyfikujesz. Klasyfikuj wyłącznie odpowiedzi, a nie tweeta. Czy odpowiedź jest homofobiczna/transfobiczna? Odpowiedzi podawaj tylko w postaci jednej cyfry: 1, jeśli odpowiedź jest homofobiczna/transfobiczna, 0, jeśli taka nie jest”. Wykorzystano wyłącznie treść oryginalnego tweeta i treść odpowiedzi. Usunięto nazwy użytkowników_użytkowniczek, linki i metadane. Amnesty International podkreśla, że kategoryzacje z użyciem dużych modeli językowych nie zawsze są stabilne i mogą się od siebie różnić w zależności od tego, którego dnia zostały przetłumaczone. Kategoryzacja przeprowadzona przez duży model językowy została wykorzystana w dalszych analizach w ograniczonym stopniu. Ilekroć miała wpływ na wnioski, Amnesty International podaje stosowne zastrzeżenia.



RYS. 1. DZIENNA LICZBA AKTYWNYCH KONT MARIONETKOWYCH W OKRESIE GROMADZENIA DANYCH



Rysunek 1. pokazuje, jak wiele kont marionetkowych było aktywnych każdego dnia. Szóstego i 26. dnia odnotowano liczne problemy techniczne. Aby zadbać o rzetelność wyników, sprawdzono, czy konta, które były w tych dniach offline, rozkładały się po równo między wszystkie cztery grupy wyróżnione na użytek badania.

24 sierpnia 2024 roku Amnesty International przesłała dostawcy usług X pytania dotyczące jego działalności biznesowej w Polsce w latach 2019–2024, ale nie otrzymała odpowiedzi. 25 czerwca 2025 roku Amnesty International poinformowała dostawcę usług o zarzutach przedstawionych w niniejszym raporcie i dała mu możliwość ustosunkowania się do nich, ale nie uzyskała odpowiedzi.

Nazwa X w raporcie dotyczy przedsiębiorstwa X Corp. (dawniej Twitter). Jest stosowana również w odniesieniu do okresu przed kwietniem 2023 roku, kiedy firma przeprowadziła rebranding. W cytatach z wywiadów nazwy „Twitter” i „X” używane są wymiennie i dotyczą zarówno platformy, jak i firmy.

3. TŁO

3.1 CEL NAGONKI: SPOŁECZNOŚĆ LGBTI W POLSCE

Osoby LGBTI od kilkudziesięciu lat zmagają się w Polsce z systemową dyskryminacją ze strony kolejnych rządów, często stosowaną w imię obrony tak zwanych tradycyjnych wartości. W latach 1985–1987 władze przeprowadziły szeroko zakrojoną tajną operację o kryptonimie Hiacynt, w ramach której zatrzymano 11 tysięcy osób podejrzanych o homoseksualizm lub kontakty ze środowiskiem osób homoseksualnych⁴. Zgromadzone wówczas potencjalnie kompromitujące informacje były określane przez Służbę Bezpieczeństwa mianem „różowych teczek”⁵. Władze nazwały operację działaniem prewencyjnym, mającym na celu walkę z pracą seksualną i „grupami przestępczymi homoseksualistów”⁶. Od czasu demokratyzacji w latach dziewięćdziesiątych prawa osób LGBTI stanowią niezwykle istotną kwestię w polskiej polityce⁷.

W okresie rządów PiS, czyli w latach 2015–2023, polskie władze podejmowały działania zmierzające do skurczenia przestrzeni działań obywatelskich: dążyły do erozji rządów prawa, atakowały prawa kobiet i osób LGBTI, tworzyły atmosferę wrogości wobec osób LGBTI i ich sojuszników⁸. Wroga, stygmatyzująca retoryka, którą posługiwali się między innymi wysocy urzędnicy państwowi, prowadziła niekiedy do aktów przemocy i dyskryminacji ze względu na prawdziwą lub postrzeganą orientację seksualną oraz tożsamość i ekspresję płciową⁹. Politycy i polityczki PiS, czyli partii rządzącej, szczególnie często posługiwali się hasłami anty-LGBTI, podsycając atmosferę wrogości i dając przyzwolenie na nienawiść do osób LGBTI¹⁰.

Specjalistka w zakresie praw osób transpłciowych i niebinarnych powiedziała Amnesty International, że po dojściu PiS do władzy wypowiedzi wrogie wobec społeczności LGBTI stały się o wiele powszechniejsze.

„Kiedy PiS zaczął rządzić, rzeczy, które [wcześniej] były nieakceptowalne... Ludzie zaczęli być z nich dumni. Głoszenie nienawiści stało się supersprawą”¹¹.

Wiele grup występujących przeciwko prawom człowieka cieszyło się coraz większą bezkarnością, gdy usprawiedliwiała przemoc i dyskryminację osób LGBTI lub wręcz wzywała do takiej przemocy bądź się jej dopuszczała¹². Amnesty International wykazała, że istnieje bezpośredni związek między erozją prawa jednostek do wolności słowa, wyrażania siebie i wolności pokojowych zgromadzeń a nękaniami, profilowaniem i zastraszaniem osób aktywistycznych LGBTI w Polsce¹³.

Ludzie piastujący wysokie stanowiska w rządzie PiS w latach 2015–2023 i inne wpływowe osoby publiczne przedstawiały społeczność LGBTI jako zagrożenie dla wartości rodzinnych, wiary katolickiej oraz porządku publicznego¹⁴. Przykładowo 17 kwietnia 2018 roku Jarosław Kaczyński, prezes Prawa i Sprawiedliwości,

⁴ Amnesty International, „Byliśmy traktowani jak przestępcy”.

⁵ Tamże.

⁶ Tamże.

⁷ The Guardian, *In Poland, the home of „LGBT-free zones”, there is hope at last for the queer community*, 1 listopada 2023, <https://www.theguardian.com/commentisfree/2023/nov/01/poland-lgbtq-new-government-law-and-justice-equality>.

⁸ Amnesty International, „Byliśmy traktowani jak przestępcy”.

⁹ Tamże.

¹⁰ Tamże.

¹¹ Amnesty International, wywiad z Julią Katą, 29 lipca 2024.

¹² Amnesty International, „Byliśmy traktowani jak przestępcy”.

¹³ Tamże.

¹⁴ ILGA Europe, *Poland: Anti-LGBTI hate timeline*, 2021, <https://www.ilga-europe.org/report/poland-anti-lgbti-hate-timeline>; Amnesty International, „Byliśmy traktowani jak przestępcy”.

oznajmił podczas kampanii wyborczej przed wyborami samorządowymi: „póki my rządymy, żadnych małżeństw homoseksualnych nie będzie”, a kraje UE muszą „otrzeźwieć”¹⁵.

W czerwcu 2020 roku, podczas spotkania z mieszkańcami Brzegu, ówczesny prezydent Andrzej Duda nazwał osoby LGBTI „ideologią”: „Próbuje się nam, proszę państwa, wmówić, że to ludzie. A to jest po prostu ideologia. Jeżeli ktoś ma jakiegokolwiek wątpliwości, czy to jest ideologia, czy nie, to niech sobie zajrzy w karty historii i zobaczy, jak wyglądało na świecie budowanie ruchu LGBT, niech zobaczy, jak wyglądało budowanie tej ideologii, jakie poglądy głosili ci, którzy ją budowali. [...] Nie po to pokolenie moich rodziców przez 40 lat walczyło, żeby wyrzucić ideologię komunistyczną ze szkół [...], żebyśmy teraz godzili się na to, żeby przyszła inna ideologia, jeszcze bardziej niszcząca dla człowieka, ideologia, która pod frazesami szacunku i tolerancji ukrywa głęboką nietolerancję i eliminację, wykluczenie wszystkich tych, którzy nie chcą jej ulec”¹⁶.

W tym samym roku Komisja Europejska zapowiedziała, że zablokuje przekazywanie środków finansowych polskim miastom i regionom, które przyjęły „uchwały anty-LGBT”. Minister Sprawiedliwości zareagował na to deklaracją, że te miasta i regiony będą mogły liczyć na wsparcie z Funduszu Sprawiedliwości przeznaczonego dla ofiar przestępstw¹⁷. Jak tłumaczyło Ministerstwo Sprawiedliwości, stały się one „ofiarami” unijnej polityki¹⁸.

W marcu 2021 roku Zbigniew Ziobro, łączący stanowiska Ministra Sprawiedliwości i prokuratora generalnego, złożył w sejmie projekt ustawy zakazującej adopcji dzieci przez pary jedнопłciowe w Polsce¹⁹.

W 2022 roku Amnesty International przedstawiła jednoznaczne dowody na to, że coraz częściej dochodziło do ataków na pokojowe zgromadzenia organizowane przez społeczność LGBTI, co stanowiło skutek kampanii nienawiści zainicjowanej przez rząd i prowadzonej zwłaszcza przez polityków polityczki PiS i Konfederacji, szczególnie w okresie poprzedzającym wybory prezydenckie w 2020 roku²⁰.

W miarę jak stygmatyzacja osób LGBTI w Polsce przybierała na sile, Marsze Równości i inne pokojowe zgromadzenia zaczęły się potykać z wrogością ze strony władz centralnych i lokalnych, a także z przemocą ze strony funkcjonariuszy organów ścigania²¹. Głośnym przykładem był Marsz Równości w Białymstoku 20 lipca 2019 roku. Policja nie zapewniła jego uczestnikom dostatecznej ochrony, co ośmieliło do ataku znacznie liczniejszy od zgromadzenia, agresywny tłum czterech tysięcy demonstrantów. W rezultacie uczestnicy zostali obrzuceni butelkami, kostką brukową i petardami, wznoszono też homofobiczne okrzyki²².

Kilka miesięcy później, 28 września 2019 roku, na Marszu Równości w Lublinie policja użyła przeciwko kontrademonstrantom armatek wodnych i gazu pieprzowego. Aresztowano kilkadziesiąt osób²³. Ujawniono później, że dwoje kontrademonstrantów miało przy sobie domowej roboty materiały wybuchowe²⁴. Amnesty International udokumentowała ponadto przypadki profilowania protestujących osób LGBTI przez policję i zatrzymywań na podstawie orientacji seksualnej i tożsamości płciowej²⁵.

Ataki na prawa osób LGBTI, a zwłaszcza na marsze równości, spotykały się z poparciem części polskich mediów i polityków polityczek. 10 października 2019 roku telewizja publiczna wyemitowała paradoksalny film *Inwazja LGBT*, w którym padło twierdzenie, że uczestnicy marszów równości są opłacani przez zagraniczne organizacje pozarządowe²⁶. Zadawano też pytania, kto i w jakim celu finansuje zgromadzenia osób LGBTI, kto finansuje organizacje LGBTI i „na czym polega inwazja LGBT”²⁷.

¹⁵ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

¹⁶ Euronews, *LGBT campaigners denounced President Duda's comments on „communism”*, 15 czerwca 2020, <https://www.euronews.com/2020/06/15/polish-president-says-lgbt-ideology-is-worse-than-communism>.

¹⁷ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

¹⁸ Tamże.

¹⁹ Equaldex, „LGBT rights in Poland”, <https://www.equaldex.com/region/poland#adoption> (dostęp: 2 lipca 2025).

²⁰ Amnesty International, *„Byliśmy traktowani jak przestępcy”*.

²¹ Tamże.

²² Tamże.

²³ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

²⁴ Tamże.

²⁵ Amnesty International, *„Byliśmy traktowani jak przestępcy”*.

²⁶ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

²⁷ Tamże.

Julia Kata, psycholożka pracująca w Fundacji Trans-Fuzja, tłumaczyła w rozmowie z Amnesty International, że upowszechnianie treści anty-LGBTI w tradycyjnych mediach dawało przyzwolenie na jawną wrogość i agresję wobec osób LGBTI:

„Słuchasz tego rodzaju komentarzy w telewizji i co myślisz? Myślisz, że można mówić takie rzeczy, że to okej. Wszyscy tak mówią w telewizji. Przecież gdyby to było niewłaściwe albo obrzydliwe, to by tego nie pokazywali, prawda? No a pokazują. Czyli to musi być okej”²⁸.

W 2022 roku Sąd Okręgowy w Warszawie orzekł, że telewizja naruszyła dobra osobiste osób LGBTI pokazanych w dokumencie, nakazał wypłatę odszkodowania i przeprosiny. W uzasadnieniu sędzia podkreślił, że *Inwazja* miała służyć „zohydzeniu środowiska LGBT” i że naruszała standardy etyki mediów²⁹.

Podczas kampanii prezydenckiej w czerwcu 2020 roku Andrzej Duda publicznie podpisał Kartę Rodziny, w której zobowiązywał się między innymi do „obrony instytucji małżeństwa”, powstrzymania adopcji dzieci przez pary jedнопłciowe i ochrony dzieci i rodzin przed „ideologią LGBT”, którą nazywał obcą i gorszą od ideologii komunistycznej. Zapowiadał też, że będzie dążył do zakazu „ideologii LGBT” w instytucjach publicznych³⁰. Podczas tej samej kampanii opowiedział się za wpisaniem zakazu zawierania małżeństw jedнопłciowych do polskiej konstytucji³¹.

Andrzej Duda twierdził, że społeczność LGBTI próbuje narzucać w Polsce prawa osób LGBTI³². W podobnym tonie wypowiadał się prezes PiS Jarosław Kaczyński w wywiadzie w kwietniu 2021 roku, w którym wspominał o „ideologii LGBT”. „Nigdzie w traktatach europejskich nie ma niczego takiego, co by nas miało skłaniać czy zmuszać, żebyśmy przyjmowali te reguły, które rzeczywiście dzisiaj na świecie zaczynają dominować i są sprzeczne ze zdrowym rozsądkiem, i ograniczają radykalnie wolność, i to wolność ogromnej ilości ludzi, którzy są terroryzowani, żeby przyjąć tego rodzaju ideologię”³³.

W sierpniu 2021 roku do sejmu wpłynął obywatelski projekt ustawy „Stop LGBT”. Po pierwszym czytaniu został on skierowany do dalszych prac³⁴. Zakładał całkowity zakaz „promowania ideologii LGBT” w przestrzeni publicznej, stanowił więc poważne zagrożenie dla praw człowieka – zwłaszcza prawa do wolności słowa i wyrażania siebie oraz wolności zgromadzeń – i ewidentnie naruszał międzynarodowe standardy³⁵. W 2024 roku projekt został złożony ponownie. Rzecznik Praw Obywatelskich zaopiniował go wówczas negatywnie i zadeklarował, że jest to niezgodna z konstytucją próba ograniczenia praw społeczności LGBTI³⁶.

3.2 „STREFY WOLNE OD LGBT”

W 2019 roku kampania nienawiści weszła w nową fazę, gdy jednostki samorządu zaczęły przyjmować uchwały i rezolucje o tak zwanych strefach wolnych od LGBT³⁷. Zasadniczo uchwały i rezolucje nie przekładały się na konkretne działania, stanowiły jednak oczywistą próbę stygmatyzacji, wykluczenia i dyskryminacji osób LGBTI. Były jawnym sygnałem, że osoby LGBTI nie są mile widziane w samorządach, które przyjęły uchwały. W 2019 roku około 100 samorządów lokalnych w Polsce ogłosiło, że są „strefami

²⁸ Amnesty International, wywiad z Julią Katą, 29 lipca 2024.

²⁹ ILGA-Europe, *Annual review of the human rights of lesbian, gay, bisexual, trans, and intersex people in Poland covering the period of January to December 2022*, 2022, <https://www.ilga-europe.org/sites/default/files/2023/poland.pdf>.

³⁰ BBC News, *Poland LGBT protests: Three charged with hanging rainbow flags off statues*, 5 sierpnia 2020, <https://www.bbc.co.uk/news/world-europe-53673411>; ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

³¹ BBC News, *Poland LGBT protests: Three charged with hanging rainbow flags off statues*.

³² BBC News, *Poland LGBT protests: Three charged with hanging rainbow flags off statues*.

³³ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

³⁴ Amnesty International, *„Byliśmy traktowani jak przestępcy”*.

³⁵ Tamże; Amnesty International, *Pride „under attack” from new bill in parliament*, 28 października 2021, <https://www.amnesty.org/en/latest/news/2021/10/poland-pride-under-attack-from-a-new-bill-in-parliament>.

³⁶ Polska Agencja Prasowa, *Ombudsperson strikes down anti-LGBT bill*, 6 lutego 2024, <https://www.pap.pl/en/news/ombudsperson-strikes-down-anti-lgbt-bill>

³⁷ ILGA Europe, *Poland: Anti-LGBTI hate timeline*; Human Rights Watch, *Poland: rule of law erosion harms women, LGBT people*, 15 grudnia 2022, <https://www.hrw.org/news/2022/12/15/poland-rule-law-erosion-harms-women-lgbt-people>.

wolnymi od LGBT” albo zakazało „ideologii LGBT”³⁸. Na początku 2020 roku „strefy wolne od LGBT” obejmowały już około jednej trzeciej terytorium kraju³⁹.

W raporcie Zgromadzenia Parlamentarnego Rady Europy na temat przeciwdziałania nienawiści do osób LGBTI podkreślono, że istnienie stref „zaprzecza prawu osób LGBTI do życia i pozbawia je bezpiecznej przestrzeni”⁴⁰. Do grudnia 2022 roku ponad 90 samorządów regionalnych lub miejskich zadeklarowało, że są „wolne od ideologii LGBT” lub podpisało promowaną przez rząd Samorządową Kartę Praw Rodzin⁴¹. Członkowie rządu aktywnie wspierali strefy: osoby aktywistyczne informowały, że Minister Edukacji i Nauki rozsyłał listy poparcia do miast i samorządów, które przyjęły rezolucje przeciwko „ideologii LGBT”⁴². Tygodnik „Gazeta Polska” dystrybuował darmowe naklejki z napisem „Strefa wolna od LGBT” w geście poparcia dla rezolucji⁴³. Strefy wolne od LGBT zostały stanowczo potępione przez Unię Europejską. W 2021 roku Komisja Europejska wszczęła postępowanie naruszeniowe w związku z pogwałceniem przez Polskę praw podstawowych⁴⁴. Komisja uznała, że polskie władze nie odpowiedziały na skierowane do nich pytania dotyczące charakteru stref i skutków ich tworzenia. We wrześniu 2021 roku Komisja Europejska zapowiedziała, że zablokuje środki dla pięciu regionów, o ile nie odwołają one uchwał anty-LGBT. Cztery z pięciu regionów podporządkowały się temu żądaniu⁴⁵.

W 2022 roku Naczelny Sąd Administracyjny uznał, że uchwały o strefach wolnych od LGBT są aktem władczym bez właściwej podstawy prawnej i że ich „społecznym skutkiem jest naruszenie godności, czci i dobrego imienia oraz ściśle związanego z tymi dobrami życia prywatnego określonej grupy mieszkańców”⁴⁶. Na skutek orzeczenia większość uchwał została ostatecznie wycofana lub odrzucona przez inne sądy⁴⁷.

24 kwietnia 2025 roku zniesiono ostatnią strefę wolną od LGBT, po tym jak radni z Łańcuta w południowo-wschodniej Polsce zagłosowali za wycofaniem uchwały⁴⁸.

3.3. BRAK OCHRONY PRAWNEJ

Na podstawie polskiego Kodeksu karnego ścigane i sądzone są przestępstwa z nienawiści ze względu na rasę, pochodzenie etniczne, narodowość, wyznanie lub bezwyznaniowość. Kodeks nie przewiduje jednak ochrony przed przestępstwami z nienawiści motywowanymi wiekiem, niepełnosprawnością, płcią, tożsamością płciową, orientacją seksualną czy statusem społeczno-ekonomicznym⁴⁹. Według międzynarodowych i europejskich standardów praw człowieka wiek, niepełnosprawność, płeć, tożsamość płciowa, orientacja seksualna i status społeczno-ekonomiczny podlegają ochronie. Brak odpowiednich przepisów o przestępstwach z nienawiści motywowanych orientacją seksualną i tożsamością płciową oznacza, że motywacje te nie są brane pod uwagę przez organy ścigania i sądy⁵⁰.

W rezultacie braku ochrony prawnej nie ma odpowiednich mechanizmów instytucjonalnych służących zwalczaniu przestępczości o charakterze homofobicznym i transfobicznym. Władze nie gromadzą też systematycznie danych na temat homofobicznych i transfobicznych przestępstw, więc trudno szacować,

³⁸ PinkNews, *Poland abolishes last remaining „LGBT-free” zone in the country*, 29 kwietnia 2025, <https://www.thepinknews.com/2025/04/29/poland-abolishes-lgbt-free-zones>.

³⁹ LGBTQ Nation, *Poland finally repealed the country's last „LGBT-free zone”*, 28 kwietnia 2025, <https://www.lgbtqnation.com/2025/04/poland-finally-repealed-the-countrys-last-lgbt-free-zone/>

⁴⁰ Human Rights Watch, „Poland: rule of law erosion harms women, LGBT people”.

⁴¹ Tamże.

⁴² Tamże.

⁴³ ILGA Europe, *Poland: Anti-LGBTI hate timeline*.

⁴⁴ European Commission, *EU founding values: Commission starts legal action against Hungary and Poland for violations of fundamental rights of LGBTIQ people*, 15 lipca 2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3668; ILGA-Europe, „Poland: Anti-LGBTI hate timeline”; Human Rights Watch, „Poland: rule of law erosion harms women, LGBT people”.

⁴⁵ Human Rights Watch, *Poland: rule of law erosion harms women, LGBT people*.

⁴⁶ TVP World, *Poland's last „LGBT-free zone” officially abolished*, 25 kwietnia 2025, <https://tvpworld.com/86360798/poland-abolishes-last-lgbt-free-zone->; LGBTQ Nation, *Poland finally repealed the country's last „LGBT-free zone”*.

⁴⁷ TVP World, *Poland's last „LGBT-free zone” officially abolished*.

⁴⁸ PinkNews, *Poland abolishes last remaining „LGBT-free” zone in the country*.

⁴⁹ Amnesty International, *Dotknięci przez nienawiść, zapomniani przez prawo*.

⁵⁰ Tamże.

w jak dużym stopniu polska społeczność LGBTI jest zagrożona przestępstwami z nienawiści⁵¹. Osobom LGBTI, które doświadczają przemocy i innych przestępstw z nienawiści, utrudnia się bezpieczne zgłaszanie tego organom ścigania⁵². Wskutek wszystkich wymienionych tu czynników, mimo jednoznacznie udokumentowanego nasilania się atmosfery wrogości wobec osób LGBTI w Polsce i mimo międzynarodowej presji, by Polska wdrożyła odpowiednie środki zaradcze, nadal nie powstał spójny system raportowania i zwalczania przestępstw z nienawiści ani mowy nienawiści motywowanej orientacją seksualną i tożsamością płciową⁵³. W rezultacie w latach 2020–2024 Polska co roku zajmowała ostatnie miejsce wśród krajów Unii Europejskiej w rankingu badającym ochronę społeczności LGBTI⁵⁴.

Osoby LGBTI w Polsce nadal doświadczają przemocy i dyskryminacji ze względu na swoją orientację seksualną oraz tożsamość i ekspresję płciową, przy czym problem ten dotyka szczególnie osób transpłciowych i interpłciowych⁵⁵. Wiele osób LGBTI w Polsce żyje w stanie ciągłego zagrożenia. Z badania ankietowego przeprowadzonego w 2023 roku przez Agencję Praw Podstawowych Unii Europejskiej wynika, że 74% osób LGBTI w Polsce zawsze lub często unika trzymania się za ręce z osobą partnerską w miejscu publicznym, a 61% uważa, że przemoc wobec społeczności przybrała na sile od 2019 roku⁵⁶.

W 2024 roku nowy polski rząd poczynił kroki, by kategoria przestępstw z nienawiści objęła również przestępstwa motywowane orientacją seksualną, płcią, wiekiem i niepełnosprawnością⁵⁷. Projekt nowelizacji Kodeksu karnego został przyjęty przez rząd i złożony w sejmie w listopadzie 2024 roku. Pierwsze czytanie odbyło się 19 grudnia 2024 roku⁵⁸. Ministerstwo Sprawiedliwości deklarowało, że „celem projektu jest dostosowanie przepisów do konstytucyjnych i europejskich standardów”⁵⁹.

Choć należy uznać to za krok we właściwą stronę, nie obyło się bez pewnych kontrowersji. W pierwotnym projekcie proponowano, żeby ustanowić ochronę prawną również ze względu na tożsamość płciową⁶⁰. Ostatecznie jednak Ministerstwo Sprawiedliwości uznało, że ochrona przed przestępstwami z nienawiści motywowanymi płcią będzie w zupełności wystarczająca⁶¹. Decyzja ta była krytykowana przez polskie organizacje LGBTI: Lambda nazwała ją niepokojącą, a Fundacja Trans-Fuzja ostrzegała, że zmiana oznacza, iż osoby transpłciowe, czyli grupa doświadczająca skrajnego wykluczenia i szczególnie narażona na przemoc, nie zostanie objęta ochroną państwa⁶². Organizacje obywatelskie wyrażały też obawy, że istotną wadą nowelizacji jest brak definicji przestępstwa z nienawiści⁶³.

Jeśli parlament przyjmie nowelizację, prezydent może ją podpisać, zawetować lub odesłać do Trybunału Konstytucyjnego⁶⁴.

3.4 ZMIANA SYTUACJI POLITYCZNEJ

Po wyborach przeprowadzonych w październiku 2023 roku utworzono nowy rząd, w skład którego weszły Koalicja Obywatelska, Trzecia Droga i Lewica. Premierem został Donald Tusk. Część społeczności LGBTI

⁵¹ Tamże.

⁵² Amnesty International, „Byliśmy traktowani jak przestępcy”.

⁵³ Amnesty International, „Byliśmy traktowani jak przestępcy”.

⁵⁴ The Guardian, *In Poland, the home of „LGBT-free zones”, there is hope at last for the queer community*; ILGA-Europe, Rainbow Map 2024, 15 kwietnia 2024, <https://www.ilga-europe.org/report/rainbow-map-2024>.

⁵⁵ FRA, *LGBTIQ equality at a crossroads: progress and challenges: EU LGBTIQ survey III*, 2024, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2024-lgbtiq-equality_en.pdf.

⁵⁶ Tamże.

⁵⁷ Notes from Poland, *Polish government approves criminalisation of anti-LGBTI hate speech*, 28 listopada 2024, <https://notesfrompoland.com/2024/11/28/polish-government-approves-criminalisation-of-anti-lgbt-hate-speech/>; Brussels Signal, *Polish government approves criminalisation of anti-LGBT hate speech*, 3 grudnia 2024, <https://brusselssignal.eu/2024/12/polish-government-approves-criminalisation-of-anti-lgbt-hate-speech>.

⁵⁸ KPH, *Amendment to the criminal code on hate crimes and hate speech is about ensuring everyone's safety – NGOs claim*, 19 grudnia 2024, <https://kph.org.pl/en/amendment-to-the-criminal-code-on-hate-crimes-and-hate-speech-is-about-ensuring-everyones-safety-ngos-claim>.

⁵⁹ Notes from Poland, *Polish government approves criminalisation of anti-LGBTI hate speech*.

⁶⁰ Tamże.

⁶¹ Tamże.

⁶² Tamże.

⁶³ KPH, *Amendment to the criminal code on hate crimes and hate speech is about ensuring everyone's safety – NGOs claim*.

⁶⁴ Notes from Poland, *Polish government approves criminalisation of anti-LGBTI hate speech*.

wyrażała ostrożne nadzieje na poprawę sytuacji⁶⁵. Pierwszy raz w historii Polski utworzono stanowisko Ministra do spraw Równości⁶⁶. Ponadto badania sondażowe pokazywały jednoznacznie, że polskie społeczeństwo jest coraz bardziej progresywne w kwestii praw osób LGBTI. Według sondaży pracowni Ipsos z 2022 roku dwie trzecie ankietowanych popierało małżeństwa jedнопłciowe lub związki partnerskie, a 60% deklaroowało, że strefy wolne od LGBT powinny zostać zniesione⁶⁷.

Utworzenie nowego rządu było dla społeczności LGBTI swego rodzaju punktem zwrotnym. Padły obietnice zmian mających na celu lepsze zabezpieczenie praw osób LGBTI i przystosowanie części przepisów obowiązujących w Polsce do międzynarodowych standardów⁶⁸. W grudniu 2023 roku Ministerstwo Sprawiedliwości wykonało ważny gest, wystosowało bowiem publiczne przeprosiny do społeczności LGBTI za to, jak traktowali ją wcześniej przedstawiciele państwa i mediów publicznych⁶⁹.

Istnieją jednak zasadnicze obawy dotyczące tego, jak szybko wprowadzone zostaną istotne zmiany prawa⁷⁰. Poza tym najbardziej progresywne zmiany mogą się okazać niemożliwe, ponieważ w koalicji rządzącej nie ma jednomyślności co do praw społeczności LGBTI⁷¹.

Niemniej rząd poczynił kroki, żeby wypełnić część swoich obietnic. W październiku 2024 roku ogłosił projekt ustawy o związkach partnerskich⁷². Projekt zakładał, że osobom w związkach partnerskich będą przysługiwały prawo do dziedziczenia po zmarłym partnerze, a także dostęp do informacji medycznych. Nie uwzględniono w nim natomiast prawa do adopcji ani przysposobienia dzieci. Uważa się, że jest to ustępstwo mające na celu pozyskanie poparcia konserwatywnego Polskiego Stronnictwa Ludowego, partii wchodzącej w skład rządu Donalda Tuska⁷³. W kwietniu 2025 roku rząd ponownie znalazł się pod presją, by przyspieszyć prace legislacyjne, kiedy Europejski Trybunał Praw Człowieka orzekł, że Polska narusza Konwencję o ochronie praw człowieka i podstawowych wolności, gdyż nie uznaje związków osób tej samej płci i nie zapewnia im ochrony⁷⁴. EPTCz odnotował, że brak możliwości transkrypcji zagranicznego aktu małżeństwa osób tej samej płci do polskiego rejestru narusza prawa jednostki, i podkreślił, że Polska jest zobowiązana przyjmując przepisy, które pozwolą zmienić ten stan rzeczy⁷⁵.

Dokonały się też inne znaczące zmiany. W marcu 2025 roku Sąd Najwyższy wydał przełomowy wyrok dotyczący praw osób transpłciowych⁷⁶. Wcześniej osoby transpłciowe – zarówno dzieci, jak i dorośli – ubiegające się o korektę swoich danych w aktach stanu cywilnego musiały podać swoich rodziców do sądu, przyjmowano bowiem zasadę, że w każdym postępowaniu cywilnym muszą być dwie przeciwne strony⁷⁷.

⁶⁵ The Guardian, *In Poland, the home of „LGBT-free zones”, there is hope at last for the queer community*.

⁶⁶ Reuters, *Rights court rules Poland should recognise same-sex partnerships*, 12 grudnia 2023, <https://www.reuters.com/world/europe/rights-court-rules-poland-should-recognise-same-sex-partnerships-2023-12-12>.

⁶⁷ ILGA-Europe, *Annual review of the human rights of lesbian, gay, bisexual, trans and intersex people in Poland covering the period of January to December 2022*; The Guardian, *In Poland, the home of „LGBT-free zones”, there is hope at last for the queer community*; United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland, 18–29 listopada 2024, *End of mission statement*, <https://www.ohchr.org/sites/default/files/documents/issues/sexualorientation/statements/2024-11-29-preliminary-observations-ie-sogi-visit-poland.pdf>.

⁶⁸ Euronews, *Queer in Poland: when can the LGBTQ+ community expect equal rights?*, 5 kwietnia 2024, <https://www.euronews.com/2024/04/05/queer-in-poland-when-can-the-lgbtq-community-expect-equal-rights>; United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland, 18–29 listopada 2024, *End of mission statement*.

⁶⁹ United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland, 18–29 listopada 2024, *End of mission statement*.

⁷⁰ Euronews, *Queer in Poland: when can the LGBTQ+ community expect equal rights?*; Reuters, *Poland publishes civil partnership bill in boost for LGBT couples*, 18 października 2024, <https://www.reuters.com/world/europe/poland-publishes-civil-partnership-bill-boost-lgbt-couples-2024-10-18>.

⁷¹ Euronews, *Queer in Poland: when can the LGBTQ+ community expect equal rights?*.

⁷² Reuters, *Poland publishes civil partnership bill in boost for LGBT couples*.

⁷³ Reuters, *Poland publishes civil partnership bill in boost for LGBT couples*.

⁷⁴ Brussels Signal, *ECHR orders Poland to recognise same-sex partnerships*, 28 kwietnia 2025, <https://brusselssignal.eu/2025/04/echr-orders-poland-to-recognise-same-sex-partnerships/>.

⁷⁵ Brussels Signal, *ECHR orders Poland to recognise same-sex partnerships*.

⁷⁶ Human Rights Watch, *Landmark Ruling on Trans Rights in Poland*, 13 marca 2025, <https://www.hrw.org/news/2025/03/13/landmark-ruling-trans-rights-poland>.

⁷⁷ Tamże.

Powodowało to niepotrzebne komplikacje, koszty i stres⁷⁸. Orzeczenie wyeliminowało konieczność pozywania rodziców⁷⁹.

Mimo tych postępów prawa osób LGBTI, a zwłaszcza osób transpłciowych, pozostają w Polsce kwestią sporną⁸⁰. Co prawda PiS już nie rządzi, ale szkodliwa retoryka nadal negatywnie wpływa na proces prawnego uzgadniania płci i na prawa osób LGBTI. Aktywiści_aktywistki i eksperci_ekspertki określają to mianem „odgórnej polaryzacji”: wpływowi politycy i polityczki wzmacniają uprzedzenia i podsycają nienawiść do osób LGBTI, żeby dzielić polskie społeczeństwo⁸¹.

Niezależny Ekspert ONZ ds. ochrony przed przemocą i dyskryminacją ze względu na orientację seksualną i tożsamość płciową orzekł po wizycie w Polsce w 2024 roku:

„Lata wrogiej retoryki i dyskryminacyjnych praktyk pozostawiły trwałe ślady”⁸².

3.5 STATUS PLATFORMY X NA POLSKIEJ SCENIE POLITYCZNO-MEDIALNEJ

Na początku 2025 roku w Polsce było 34,5 miliona użytkowników_użytkowniczek internetu, co odpowiada prawie 90% populacji, i 29 milionów użytkowników_użytkowniczek mediów społecznościowych, co odpowiada 75,6% populacji⁸³.

Z danych publikowanych w materiałach reklamowych X wiadomo, że na początku 2025 roku platforma miała w Polsce 5,33 miliona użytkowników_użytkowniczek, co odpowiada 13,9% populacji⁸⁴. Potencjalny zasięg reklamowy platformy w Polsce spadł między październikiem 2024 roku a styczniem 2025 roku o 468 tysięcy użytkowników_użytkowniczek (o 8,4%)⁸⁵. Podobny trend spadkowy platforma notuje na całym świecie⁸⁶. Trzeba jednak podkreślić, że dane dotyczące zasięgu reklam nie informują o liczbie aktywnych użytkowników_użytkowniczek. Ta ostatnia nie jest podawana do wiadomości publicznej. Niewykluczone, że istnieje znacząca różnica między liczbą odbiorców_odbiornych reklam zamieszczanych na platformie X a liczbą aktywnych użytkowników_użytkowniczek⁸⁷.

X jest uważany za istotne źródło informacji oraz platformę, na której toczą się dyskusje w ważnych kwestiach politycznych. Mateusz Kaczmarek, członek zarządu Grupy Stonewall, polskiej organizacji LGBTI, powiedział Amnesty International:

„To najszybszy sposób na zdobycie nowych informacji. Kiedy coś się dzieje, łatwiej dowiedzieć się o tym z Twittera niż z Google’a czy z całej reszty internetu”⁸⁸.

⁷⁸ Tamże.

⁷⁹ Tamże.

⁸⁰ Tamże.

⁸¹ Euronews, *Queer in Poland: when can the LGBTQ+ community expect equal rights?*; Human Rights Watch, *Landmark Ruling on Trans Rights in Poland*.

⁸² United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland (18–29 listopada 2024), *End of mission statement*, akapit 12.

⁸³ DataReportal, Digital 2025: Poland, 3 marca 2025, <https://datareportal.com/reports/digital-2025-poland>.

⁸⁴ Tamże.

⁸⁵ Tamże.

⁸⁶ DataReportal, X Users, Stats, Data & Trends for 2025, 12 marca 2025, <https://datareportal.com/essential-x-stats>.

⁸⁷ DataReportal, Digital 2025: Poland.

⁸⁸ Amnesty International, wywiad z Mateuszem Kaczmarkiem, 28 lipca 2024.

4. UWARUNKOWANIA PRAWNE

4.1. BIZNES I STANDARDY PRAW CZŁOWIEKA

Zgodnie z prawem międzynarodowym państwa mają obowiązek szanować, chronić i wypełniać prawa człowieka. Muszą też zapobiegać ich naruszeniom ze strony podmiotów prywatnych, między innymi poprzez ustanawianie odpowiednich regulacji dla firm i innych podmiotów gospodarczych, a także wdrażanie skutecznych środków naprawczych, gdy podmioty gospodarcze na terenie danego państwa łamią prawa człowieka lub się do tego przyczyniają. Aby temu zapobiegać, niezbędne jest podejmowanie odpowiednich kroków: prowadzenie dochodzeń, wymierzanie kar i naprawianie krzywd. Wymaga to prowadzenia skutecznej polityki w wielu obszarach, odpowiednich przepisów, regulacji oraz sprawnie działającego sądownictwa. Wszystkie wymienione tu obowiązki wynikają z traktatów ratyfikowanych przez państwa oraz z innych międzynarodowych standardów. Państwa nie wypełniają swoich zobowiązań, jeśli same naruszają prawa człowieka lub jeśli nie czynią odpowiednich kroków, żeby zapobiegać łamaniu praw człowieka przez podmioty prywatne – nie prowadzą dochodzeń w tym zakresie, nie wymierzają kar i nie naprawiają krzywd.

Wytyczne dotyczące biznesu i praw człowieka przyjęte przez Organizację Narodów Zjednoczonych jasno wskazują, że państwo ma obowiązek chronić obywateli i obywatelki przed naruszeniami praw człowieka związanymi z działalnością gospodarczą⁸⁹. Ponadto każda firma jest zobowiązana respektować prawa człowieka bez względu na swoją branżę, wielkość i to, w której części świata prowadzi działalność⁹⁰. Obowiązek ten jest niezależny od powinności państwa w zakresie praw człowieka i nadrzędny wobec obowiązku przestrzegania krajowych przepisów i regulacji dotyczących praw człowieka⁹¹.

Wytyczne OECD dla przedsiębiorstw wielonarodowych dotyczące odpowiedzialnego prowadzenia działalności biznesowej (dalej: *Wytyczne OECD*) zawierają zalecenia dla rządów i firm, są ponadto wiążące dla państw członkowskich takich jak Polska⁹². *Wytyczne OECD dotyczące należytej staranności w odpowiedzialnym prowadzeniu działalności biznesowej* (dalej: *Wytyczne OECD dotyczące należytej staranności*) zostały stworzone, żeby zapewnić przedsiębiorstwom praktyczne wsparcie, pomocne we wdrażaniu *Wytycznych OECD*. *Wytyczne OECD dotyczące należytej staranności* uwzględniają między innymi kwestię przyczyniania się przez przedsiębiorstwa do powstania negatywnych skutków, bezpośrednio powiązanych z ich działaniami. O przyczynianiu się przez dane przedsiębiorstwo do negatywnego skutku można mówić wówczas, jeśli wniosło ono „istotny wkład, oznaczający działanie powodujące, ułatwiające lub zachęcające inny podmiot do spowodowania negatywnego skutku”⁹³. Nie obejmuje to „drobnych ani nieznaczących działań”⁹⁴.

Ponadto *Wytyczne dotyczące biznesu i praw człowieka* dyktują, że przedsiębiorstwa powinny dochowywać należytej staranności, czyli prowadzić proaktywny proces mający na celu zidentyfikowanie faktycznych i potencjalnych naruszeń praw człowieka, których mogą się dopuszczać. Celem procesu należytej staranności jest unikanie i łagodzenie negatywnych skutków. Jeśli przedsiębiorstwo ustali, że narusza prawa

⁸⁹ *Wytyczne dotyczące biznesu i praw człowieka: Wdrażanie dokumentu ramowego ONZ „Chronić, szanować i naprawiać”, zasada 1.*

⁹⁰ Obowiązek ten został ustanowiony przez Radę Praw Człowieka ONZ 16 czerwca 2011 roku, kiedy przyjęła *Wytyczne dotyczące biznesu i praw człowieka* i 25 maja 2011 roku, kiedy rządy 42 państw wdrażających *Deklarację w sprawie inwestycji międzynarodowych oraz przedsiębiorstw wielonarodowych* OECD jednogłośnie przyjęło nowe wytyczne OECD dla przedsiębiorstw wielonarodowych. Zob. Human Rights and Transnational Corporations and other Business Enterprises, Human Rights Council, Resolution 17/4, 6 lipca 2011, UN Doc. A/HRC/RES/17/4; OECD, OECD Guidelines for Multinational Enterprises, 2011, <https://www.oecd.org/en/topics/responsible-business-conduct.html>.

⁹¹ *Wytyczne dotyczące biznesu i praw człowieka*, zasada 11. i komentarz.

⁹² *Wytyczne OECD dla przedsiębiorstw wielonarodowych dotyczące odpowiedzialnego prowadzenia działalności biznesowej*, 2023, https://www.oecd.org/content/dam/oecd/pl/publications/reports/2023/06/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct_a0b49990/9fe6d8dc-pl.pdf.

⁹³ *Wytyczne OECD dla przedsiębiorstw wielonarodowych dotyczące odpowiedzialnego prowadzenia działalności biznesowej*, s. 20.

⁹⁴ Tamże, s. 20.

człowieka lub że jego działania mogą do tego prowadzić, musi natychmiast zareagować, żeby negatywne skutki przestały występować lub żeby ograniczyć ryzyko ich wystąpienia⁹⁵.

Konieczne jest także monitorowanie skuteczności działań podejmowanych przez firmę w celu rozwiązania problemu ewentualnych naruszeń⁹⁶. „Monitorowanie to powinno: (a) opierać się na odpowiednich wskaźnikach ilościowych i jakościowych; (b) korzystać z informacji zwrotnych ze źródeł wewnętrznych i zewnętrznych, w tym od interesariuszy dotkniętych tym oddziaływaniem”⁹⁷.

4.2 PRAWA OSÓB LGBTI W KONTEKŚCIE BIZNESU

W 2017 roku biuro Wysokiego Komisarza Narodów Zjednoczonych ds. Praw Człowieka przedstawiło standardy postępowania dla przedsiębiorstw dotyczące przeciwdziałania dyskryminacji osób LGBTI⁹⁸. Przypomniano, że firmy są zobowiązane przestrzegać praw człowieka, a ponadto sformułowano konkretne zalecenia w zakresie tego, jak mają dbać o prawa osób LGBTI⁹⁹. Wśród standardów znalazł się też następujący zapis: „Jeśli udokumentowano powszechne naruszenia praw osób LGBTI, w tym w krajach, w których obowiązują dyskryminacyjne przepisy i stosuje się dyskryminacyjne praktyki, przedsiębiorstwa muszą dokładać jeszcze większej należytej staranności, by zagwarantować, że będą dbały o prawa osób LGBTI”¹⁰⁰.

W 2024 roku w raporcie Grupy Roboczej ONZ ds. Biznesu i Praw Człowieka powtórzono, że państwa i firmy muszą dokładać starań, żeby zapobiegać negatywnym skutkom działalności biznesowej godzącym w prawa osób LGBTI. „Osoby ze społeczności LGBTI+ są narażone na różnorodne ryzyka, a strukturalna i interseksjonalna dyskryminacja, których doświadczają, jest często niedostatecznie dobrze rozumiana przez państwa i przedsiębiorstwa, co prowadzi do nieadekwatnych reakcji”¹⁰¹.

Grupa robocza przypomniła, że osoby LGBTI są narażone na różne formy dyskryminacji i ataków, w tym na stygmatyzację w mediach społecznościowych¹⁰². W raporcie napisano, że platformy cyfrowe często nie wypełniają swoich obowiązków wobec społeczności LGBTI i przestrzegania praw człowieka:

„Platformy cyfrowe są wykorzystywane do rozpowszechniania szkodliwych treści i nawoływania do przemocy. Co gorsza, platformy te stosują zautomatyzowane systemy moderowania treści, które nie biorą pod uwagę ani problemów związanych z prawami człowieka, ani różnorodności językowej. Algorytmy platform społecznościowych nie potrafią identyfikować mowy nienawiści w ramach różnych dialektów, przez co wykorzystywane przez firmy mechanizmy rozpatrywania skarg stają się nieskuteczne. Aby zagwarantować, że mechanizmy rozpatrywania skarg pozwolą skutecznie rozwiązać problemy, z którymi regularnie zmagają się użytkownicy_użytkowniczki, niezbędne będą odpowiednie konsultacje z osobami LGBTI+”¹⁰³.

Raport Grupy Roboczej zawiera zalecenia dla przedsiębiorstw w zakresie wypełniania obowiązku przestrzegania praw osób LGBTI, między innymi zalecenia dotyczące procesu należytej staranności uwzględniającego kwestie płci i tożsamości płciowej oraz dialogu ze społecznością LGBTI¹⁰⁴.

⁹⁵ „Wytyczne dotyczące biznesu i praw człowieka”, komentarz do zasady 19.

⁹⁶ „Wytyczne dotyczące biznesu i praw człowieka”, zasada 20.

⁹⁷ „Wytyczne dotyczące biznesu i praw człowieka”, komentarz do zasady 21.

⁹⁸ OHCHR, *Tackling Discrimination against Lesbian, Gay, Bi, Trans & Intersex People: Standards of Conduct for Business*, 2017, https://www.unfe.org/sites/default/files/documents/UN-Standards-of-Conduct_0.pdf.

⁹⁹ Tamże.

¹⁰⁰ Tamże.

¹⁰¹ OHCHR, *UN Working Group calls for urgent efforts to address violations and abuses of the rights of LGBTI+ persons in business contexts*, 1 listopada 2024, <https://www.ohchr.org/en/press-releases/2024/11/un-working-group-calls-urgent-efforts-address-violations-and-abuses-rights>.

¹⁰² Tamże.

¹⁰³ Working Group on the Issue of Human Rights and Transnational Business Enterprises, *Protecting and respecting the rights of lesbian, gay, bisexual, transgender and intersex persons in the context of business activities: fulfilling obligations and responsibilities under the UN Guiding Principles on Business and Human Rights*, 18 lipca 2024, UN Doc. A/79/178, s. 22.

¹⁰⁴ OHCHR, *UN Working Group calls for urgent efforts to address violations and abuses of the rights of LGBTI+ persons in business contexts*.

4.3 NALEŻYTA STARANNOŚĆ W ZAKRESIE PRZESTRZEGANIA PRAW CZŁOWIEKA A FIRMY Z BRANŻY TECHNOLOGICZNEJ

Wytyczne dotyczące biznesu i praw człowieka prezentują istotne standardy, których powinny przestrzegać przedsiębiorstwa, między innymi standard procesów należytej staranności w posługiwaniu się technologiami opartymi na algorytmach, czyli na przykład narzędzi cyfrowych służących do moderowania treści.

W 2020 roku Wysoki Komisarz Narodów Zjednoczonych ds. Praw Człowieka przypomniał, że wytyczne mają zastosowanie również w przypadku przedsiębiorstw z branży technologicznej:

„W wytycznych zaprezentowano zasady, którymi powinny się kierować wszystkie firmy – bez względu na branżę, wielkość, strukturę czy kontekst działalności – w celu identyfikowania ryzyka i podejmowania działań, które pozwolą temu ryzyku zapobiec lub je ograniczyć. Oczekuje się, że firmy z branży technologicznej będą dokładały starań, żeby przewidywać i nie dopuszczać do negatywnych efektów mogących stanowić konsekwencję korzystania z ich produktów i usług”¹⁰⁵.

Wysoki Komisarz Narodów Zjednoczonych ds. Praw Człowieka podkreśla, że procesy należytej staranności w branży technologicznej muszą obejmować radzenie sobie z sytuacjami, w których „praktyki i decyzje wynikające z modelu biznesowego powodują lub zwiększają ryzyko naruszeń praw człowieka”. W takich sytuacjach konieczna jest analiza obejmująca konkretne ryzyko naruszenia praw człowieka związane z danymi produktami i usługami, użytkownikami/użytkowniczkami końcowymi i kontekstem użytkowania¹⁰⁶. Ponadto Specjalny Sprawozdawca ONZ do spraw Wolności Słowa przygotował „standardy dla systemów sztucznej inteligencji”, w których jednoznacznie obliuguje się przedsiębiorstwa do tego, by ich „standardy działania, projekty systemów oraz reguły opierały się na uniwersalnych zasadach dotyczących praw człowieka”¹⁰⁷.

W 2021 roku Wysoki Komisarz Narodów Zjednoczonych ds. Praw Człowieka przedstawił rekomendacje dla przedsiębiorstw w zakresie oceny ryzyka towarzyszącego posługiwaniu się sztuczną inteligencją (AI), obejmujące między innymi¹⁰⁸:

- ❑ systematyczne przeprowadzanie procesu należytej staranności w zakresie praw człowieka na każdym etapie cyklu systemu AI, który projektują, rozwijają, stosują, sprzedają, pozyskują lub wykorzystują. Kluczowym elementem tego procesu powinny być regularne, całościowe analizy dotyczące tego, jak system wpływa na prawa człowieka;
- ❑ zasadnicze zwiększenie przejrzystości w zakresie tego, jak przedsiębiorstwa korzystają z AI, między innymi dzięki informowaniu o tym społeczeństwa i dotkniętych jednostek, a także dzięki umożliwieniu niezależnych, zewnętrznych audytów systemów AI. Im większe ryzyko, że zastosowanie AI doprowadzi do naruszeń praw człowieka, tym więcej potrzeba przejrzystości;
- ❑ zadbanie o partycypację wszystkich interesariuszy – zwłaszcza dotkniętych jednostek i grup – w procesie rozwoju AI;
- ❑ dążenie do większej wyjaśnialności sztucznej inteligencji, między innymi dzięki przeznaczaniu odpowiednich środków finansowych i prowadzeniu badań na ten temat.

Międzynarodowe prawa człowieka podsuwają też branży technologicznej istotne wskazówki w zakresie podejścia do moderowania treści. Jeśli są wdrażane w sposób przejrzysty i konsekwentny, z istotnym wkładem ze strony użytkowników/użytkowniczek i społeczeństwa obywatelskiego, pozwalają stworzyć system, dzięki któremu zarówno państwa, jak i przedsiębiorstwa będą mogły być rozliczane przez użytkowników/użytkowniczki na całym świecie¹⁰⁹. Standardy praw człowieka pozwolą też firmom tworzyć inkluzywne środowisko uwzględniające różne potrzeby i interesy użytkowników/użytkowniczek, a zarazem zadbać o przewidywalne, spójne podstawowe standardy korzystania z danego produktu lub danej usługi¹¹⁰.

¹⁰⁵ OHCHR, *The UN Guiding Principles in the Age of Technology: A B-Tech Foundational Paper*, wrzesień 2020, <https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/introduction-ungp-age-technology.pdf>.

¹⁰⁶ OHCHR, *Addressing Business Model Related Human Rights Risks: A B-Tech Foundational Paper*, lipiec 2020, https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/B_Tech_Foundational_Paper.pdf.

¹⁰⁷ Special Rapporteur on the promotion and protection of the right to freedom of expression, Report: *Artificial Intelligence Technologies and Implications for Freedom of Expression and the Information Environment*, 29 sierpnia 2018, UN Doc. A/73/348, akapit 12.

¹⁰⁸ OHCHR, *The right to privacy in the digital age*, 15 września 2021, UN Doc. A/HRC/48/31.

¹⁰⁹ Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Report: *A Human Rights Approach to Platform Content Regulation*, 6 kwietnia 2018, UN Doc. A/HRC/38/25.

¹¹⁰ Tamże.

W 2018 roku Specjalny Sprawozdawca ONZ do spraw Wolności Słowa przedstawił propozycje działań umożliwiających moderowanie treści oparte na standardach praw człowieka¹¹¹. Wśród sugerowanych rozwiązań znalazło się: odejście od zasad użytkowania, w których dominuje dyskrejonalny nacisk na bliżej nieokreślone potrzeby „społeczności” i zobowiązanie się do prowadzenia platform w sposób, który zapewnia użytkownikom_użytkowniczkom możliwość kształtowania własnych opinii, swobodnego wyrażania siebie i uzyskiwania dostępu do wszelkiego rodzaju informacji¹¹². Na tym właśnie powinno bazować podejście do moderowania treści. Należy zagwarantować, że wszystkie działania dotyczące treści i ich moderowania będą się opierały na tych samych standardach legalności, konieczności i legitymizacji, których muszą przestrzegać państwa, kiedy kształtują regulacje dotyczące wolności słowa i wyrażania siebie¹¹³. Aby jeszcze silniej powiązać praktyki w zakresie moderowania treści z prawami człowieka, przedsiębiorstwa powinny poczynić kroki, żeby wyjaśnić i skonkretyzować zasady użytkowania swoich usług. Użytkownicy_użytkowniczki powinni być w stanie przewidzieć z dużą dozą pewności, jakiego rodzaju treści, które opublikują, będą oznaczały naruszenie zasad¹¹⁴. Firmy mogą również ujawniać dane i prezentować przykłady, dzięki czemu będzie wiadomo, jakie czynniki biorą pod uwagę, gdy oceniają, czy doszło do naruszenia zasad. W kontekście mowy nienawiści przedstawienie konkretnych przykładów i sposobów reagowania pomoże użytkownikom_użytkowniczkom lepiej zrozumieć, w jaki sposób firma podchodzi do trudnego zadania polegającego na ustaleniu, gdzie przebiega granica między nieszkodliwymi treściami a nawoływaniem do przemocy lub nienawiści oraz w jaki sposób ocenia się intencje autora treści lub prawdopodobieństwo, że w wyniku publikacji danych treści może dojść do przemocy¹¹⁵.

Global Network Initiative (GNI), w skład której wchodzi najwaźniejsi interesariusze z akademii, organizacji obywatelskich i inwestorzy, również przygotowała wytyczne dotyczące moderacji treści oparte na standardach praw człowieka, skupiając się zwłaszcza na działaniach państwa dotyczących szkodliwych treści online¹¹⁶. W wytycznych podkreślono, że wszystkie procesy legislacyjne, które dotyczą tych kwestii, muszą przebiegać w sposób jawny, otwarty i niekonfrontacyjny. Ponadto przygotowując przepisy, należy korzystać z różnorodnych zasobów wiedzy eksperckiej, aby efekt końcowy był dobrze przemyślany i oparty na dowodach¹¹⁷. Jest szczególnie ważne, by władze państwowe poświęciły odpowiednio dużo czasu na przemyślenie i zrozumienie konieczności podejmowania działań wynikających z międzynarodowych zobowiązań w zakresie praw człowieka¹¹⁸. Chociaż platformy i dostawcy usług również ponoszą odpowiedzialność i odgrywają istotną rolę w przeciwdziałaniu naruszeniom praw człowieka, do których dochodzi online, państwo nie może zrzucić na nie całej odpowiedzialności prawnej za generowanie przez użytkowników_użytkowniczki nielegalnych treści¹¹⁹.

W wytycznych GNI kładzie się nacisk na transparencję, środki naprawcze i rozliczalność, które muszą stanowić nieodzowny element legislacji dotyczącej moderowania treści. Konieczne jest też przygotowanie przepisów w taki sposób, żeby umożliwiały skuteczną regulację produktów i usług, w przypadku których ryzyko naruszeń praw człowieka jest największe. Niezbędne jest tu między innymi wskazanie wyjątków i odpowiednich zabezpieczeń¹²⁰. W wytycznych podkreślono, że treści zidentyfikowane przez zautomatyzowane narzędzia jako naruszające zasady muszą być następnie dodatkowo zweryfikowane przez moderatorów. Ponadto trzeba zadbać o zagwarantowanie użytkownikom_użytkowniczkom prawa do skutecznych procedur odwoławczych w razie usunięcia treści lub ograniczenia ich widoczności¹²¹. Podejmując decyzje w zakresie moderowania treści, należy szukać równowagi między prawem do wolności słowa i wyrażania siebie a innymi prawami, na przykład prawem do ochrony przed dyskryminacją. Dlatego w wytycznych GNI zaznaczono, że przygotowanie szczególnych środków ochrony, obejmujących grupy najbardziej narażone na przemoc lub dyskryminację, pomoże zagwarantować, że przepisy spełnią swój cel¹²².

¹¹¹ Tamże.

¹¹² Tamże.

¹¹³ Tamże, akapit 44; GNI, *Content Regulation and Human Rights*, październik 2020, <https://globalnetworkinitiative.org/wp-content/uploads/GNI-Content-Regulation-HR-Policy-Brief.pdf>.

¹¹⁴ Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *Report: A Human Rights Approach to Platform Content Regulation*.

¹¹⁵ Tamże.

¹¹⁶ GNI, *Content Regulation and Human Rights*.

¹¹⁷ Tamże.

¹¹⁸ Tamże.

¹¹⁹ Tamże.

¹²⁰ Tamże.

¹²¹ Tamże.

¹²² Tamże.

4.4 ODPOWIEDZIALNOŚĆ PRZEDSIĘBIORSTW W ZAKRESIE ŚRODKÓW NAPRAWCZYCH

Środki naprawcze to niezwykle ważny filar systemu zwalczania naruszeń praw człowieka przez biznes. W *Wytycznych dotyczących biznesu i praw człowieka* przyjętych przez ONZ wskazano: „Jeżeli przedsiębiorstwo stwierdzi, że spowodowało negatywny wpływ (na prawa człowieka) lub przyczyniło się do takiego wpływu, powinno naprawić wyrządzoną szkodę poprzez przeprowadzenie cieszącego się zaufaniem zainteresowanych, rzetelnego procesu naprawczego lub poprzez współpracę w jego przeprowadzeniu”¹²³. Potencjalny negatywny wpływ – lub ryzyko takiego wpływu – wymaga działań, które zapobiegają szkodom lub ograniczą ryzyko tak bardzo, jak to możliwe. Przedsiębiorstwa nie są w stanie wypełniać swoich obowiązków w zakresie praw człowieka, jeśli w razie naruszenia tych praw nie podejmują odpowiednich działań naprawczych¹²⁴.

4.5 ZOBOWIĄZANIA WYNIKAJĄCE Z AKTU O USŁUGACH CYFROWYCH (DSA)

Unia Europejska czyniła starania, żeby uregulować działalność platform cyfrowych oraz stosowanie technologii opartych na algorytmach. W 2022 roku przyjęła Akt o usługach cyfrowych (DSA), rozporządzenie o charakterze wiążącym¹²⁵. Za egzekwowanie jego zapisów odpowiadają władze państw członkowskich. Każde państwo członkowskie ma obowiązek powołać specjalnego koordynatora lub koordynatorów ds. usług cyfrowych¹²⁶. Ponieważ zapisy DSA są wiążące, ich naruszanie pociąga za sobą ryzyko nałożenia grzywny w wysokości do 6% rocznego światowego obrotu danego dostawcy usług. Koordynatorzy ds. usług cyfrowych są ponadto uprawnieni do przeprowadzania kontroli u dostawców usług. Mają prawo żądać od nich wyjaśnień dotyczących algorytmów, przetwarzania danych i praktyk biznesowych¹²⁷.

DSA zobowiązuje platformy społecznościowe do przejrzystości w zakresie moderowania treści. Dostawcy usług powinni podawać i aktualizować informacje na temat wszelkich polityk, procedur, środków i narzędzi wykorzystywanych do celów moderowania treści¹²⁸. Zobowiązano ich też do stosowania skutecznych środków chroniących użytkowników „użytkowniczo”, takich jak umożliwienie odwołania się od decyzji o usunięciu lub zmniejszeniu widoczności określonych treści. W każdym takim przypadku platforma ma obowiązek wyjaśnić użytkownikowi „użytkowniczo”, dlaczego publikowane przez niego/ną treści zostały usunięte lub ograniczone ich widoczność¹²⁹.

Artykuł 34(1) DSA ustanawia obowiązki ciążące na bardzo dużych platformach internetowych (VLOP) związane z oceną i ograniczaniem ryzyka systemowego wynikające z projektu lub funkcjonowania ich usługi i powiązanych z nią systemów, w tym systemów algorytmicznych, lub z korzystania z ich usług¹³⁰.

Artykuł 34(1)(b) precyzuje, że ocena ryzyka musi uwzględniać specyfikę usług VLOP i być proporcjonalna do ryzyka systemowego, biorąc pod uwagę jego wagę i prawdopodobieństwo, a także obejmować:

„ryzyko wystąpienia faktycznych lub przewidywalnych negatywnych skutków dla wykonywania praw podstawowych, w szczególności prawa podstawowego do godności ludzkiej zapisanego w art. 1 Karty, do poszanowania życia prywatnego i rodzinnego zapisanego w art. 7 Karty, do ochrony danych osobowych zapisanego w art. 8 Karty, do wolności wypowiedzi i informacji, w tym wolności i pluralizmu mediów,

¹²³ *Wytyczne dotyczące biznesu i praw człowieka: wdrażanie dokumentu ramowego ONZ „Chronić, szanować i naprawiać”, zasada 22.*

¹²⁴ OHCHR, *Frequently asked questions about the UN Guiding Principles on Business and Human Rights*, 2014, <https://www.ohchr.org/en/publications/special-issue-publications/frequently-asked-questions-about-guiding-principles>, s. 36.

¹²⁵ Amnesty International, *What the EU's Digital Services Act Means for Human Rights* (Index: POL 30/5830/2022), 7 lipca 2022, <https://www.amnesty.org/en/documents/pol30/5830/2022/en>.

¹²⁶ Tamże.

¹²⁷ Tamże.

¹²⁸ European Commission, „Questions and answers on the Digital Services Act”, 23 lutego 2024, https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348.

¹²⁹ Tamże.

¹³⁰ Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych), Artykuł 43(1), <https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX:32022R2065>.

zapisanego w art. 11 Karty, do niedyskryminacji zapisanego w art. 21 Karty, do przestrzegania praw dziecka zapisanego w art. 24 Karty i do wysokiej jakości ochrony konsumentów zapisanego w art. 38 Karty”.

Podobnie jak w międzynarodowych standardach przedstawionych powyżej DSA wymaga, by w procesy należytej staranności prowadzone przez dostawców platform społecznościowych włączyć odpowiednich interesariuszy, gdyż pozwoli to skuteczniej identyfikować i ograniczać ryzyko systemowe. Wymaga się też, aby dostawcy VLOP sprawdzali swoje założenia z grupami, których w największym stopniu dotyczą ryzyko i podejmowane środki¹³¹.

Dostawcy VLOP mają obowiązek wprowadzać rozsądne, proporcjonalne i skuteczne środki zmniejszające ryzyko, dopasowane do konkretnego ryzyka systemowego ze szczególnym uwzględnieniem wpływu takich środków na prawa podstawowe¹³². Środki mogą obejmować: dostosowanie projektu, cech lub funkcjonowania ich usług, w tym dostosowanie interfejsów internetowych oraz procesów moderowania treści, a także testowanie i dostosowywanie systemów algorytmicznych, w tym systemów rekomendacji¹³³.

Dostawcy VLOP co najmniej raz w roku poddają się niezależnym audytom¹³⁴. Obowiązki ustanowione przez DSA dotyczą wybranych VLOP (w tym X) od sierpnia 2023 roku. Wszystkie platformy cyfrowe zostały nimi objęte w lutym 2024 roku¹³⁵.

4.6 PRAWA DO ŻYCIA WOLNEGO OD PRZEMOCY ZE WZGLĘDU NA PŁEĆ

Prawo międzynarodowe i standardy praw człowieka zobowiązują państwa do chronienia prawa obywateli_obywaterek do życia wolnego od przemocy ze względu na płeć (GBV). GBV obejmuje wiele rodzajów przemocy, między innymi fizyczną, seksualną i psychiczną, pogrozki, znęcanie się i przymus wynikające z nierówności płci, asymetrii władzy oraz szkodliwych stereotypów płciowych i szkodliwych norm społecznych. GBV to zjawisko, które nieproporcjonalnie częściej dotyczy kobiet i dziewczynek. Dotyka również inne osoby w zależności od ich prawdziwej lub postrzeganej płci, orientacji seksualnej, tożsamości i ekspresji płciowej. GBV stanowi formę dyskryminacji. W niektórych przypadkach sprowadza się do tortur lub innych form maltretowania. Definicja GBV obejmuje również przemoc „dokonującą się online i w innych środowiskach cyfrowych”¹³⁶.

Institucje i mechanizmy ochrony praw człowieka ONZ w coraz większym stopniu skupiają się na dyskryminacji ze względu na orientację seksualną, tożsamość płciową, ekspresję płciową i cechy płciowe (SOGIESC), która odgrywa zasadniczą rolę w upowszechnianiu GBV, w tym przemocy ze względu na płeć wspomaganą technologicznie (TFGBV). Amnesty International definiuje TFGBV jako dowolny akt GBV lub groźbę takiego aktu, których dopuszcza się jedna lub więcej osób, popełnianych, wspomaganych, wzmacnianych lub nasilanych w całości lub częściowo przez technologie informacyjne lub komunikacyjne bądź media cyfrowe. TFGBV nieproporcjonalnie częściej dotyczy kobiety i dziewczynki, ale spotyka też inne osoby ze względu na ich prawdziwą lub postrzeganą płeć lub SOGIESC, co prowadzi do krzywd o charakterze fizycznym, psychicznym, ekonomicznym, społecznym i seksualnym.

W Rekomendacji Ogólnej nr 35 dotyczącej GBV wobec kobiet Komitet ds. Likwidacji Dyskryminacji Kobiet (CEDAW) powtórzył swoje wcześniejsze stanowisko, zgodnie z którym wszelkie formy dyskryminacji kobiet mają charakter intersekcjonalny i są „nierozzerwalnie związane z innymi czynnikami”, obejmującymi bycie lesbijką, kobietą biseksualną lub transpłciową bądź osobą interpłciową¹³⁷. W raporcie poświęconym dyskryminacji ze względu na SOGIESC przedstawionym Radzie Praw Człowieka ONZ ówczesny Wysoki Komisarz Narodów Zjednoczonych ds. Praw Człowieka zaliczył do GBV homofobiczne i transfobiczne

¹³¹ Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych), Motyw 90.

¹³² Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych), Artykuł 35.

¹³³ Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2022/2065 z dnia 19 października 2022 r. w sprawie jednolitego rynku usług cyfrowych oraz zmiany dyrektywy 2000/31/WE (akt o usługach cyfrowych), Artykuł 35.

¹³⁴ Amnesty International, *What the EU's Digital Services Act Means for Human Rights*.

¹³⁵ European Commission, *The Digital Services Act*, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (dostęp: 2 lipca 2025).

¹³⁶ CEDAW Committee, General Recommendation No. 35: Gender-Based Violence Against Women, Updating General Recommendation No. 19 (1992), 26 lipca 2017, UN Doc. CEDAW/C/GC/35, akapit 20.

¹³⁷ CEDAW Committee, General Recommendation No. 35, akapit 12.

ataki¹³⁸. Ponadto Zasady Yogyakarty dotyczące stosowania prawa międzynarodowego w celu przeciwdziałania dyskryminacji ze względu na SOGIESC chronią prawa osób LGBTI jako użytkowników_ użytkowników technologii informacyjnych i komunikacyjnych. Zasada 36. głosi, że osoby LGBTI mają prawo do takiej samej ochrony online i offline i że mają prawo do korzystania z technologii komunikacyjnych bez zagrożenia przemocą i dyskryminacją ze względu na SOGIESC¹³⁹.

Międzynarodowe prawa człowieka nakładają na państwa obowiązek zagwarantowania, że zarówno państwowe, jak i niepaństwowe podmioty będą szanować prawo osób LGBTI do życia wolnego od GBV, w tym TFGBV¹⁴⁰. Państwa muszą podejmować wszelkie niezbędne kroki, żeby chronić osoby narażone na GBV, w tym TFGBV¹⁴¹, prowadzić śledztwa w sprawie naruszeń, stawiać sprawców przed sądem, zapewniać ofiarom dostęp do wymiaru sprawiedliwości oraz szybkie i stosowne reparacje¹⁴². Państwa mają też zapobiegać TFGBV poprzez upowszechnianie świadomości i zapewnianie wsparcia wszystkim osobom, które doświadczyły GBV¹⁴³. Nadrzędnie ważną kwestią jest tu stosowanie podejścia intersekcjonalnego i branie pod uwagę rozmaitych sposobów, w jakie rasa, etniczność i status społeczno-ekonomiczny wpływają na doświadczenia TFGBV w różnych kontekstach¹⁴⁴.

Prawo do życia wolnego od GBV jest nierozdzielnie powiązane z innymi prawami człowieka, między innymi z prawem do prywatności, prawem do wolności słowa i wyrażania siebie, prawem do pokojowych zgromadzeń i prawem do stowarzyszania się¹⁴⁵.

Kiedy ludzie doświadczają TFGBV, nie mogą bezpiecznie uczestniczyć w dyskusjach toczonych online ani korzystać z technologii cyfrowych takich jak platformy społecznościowe¹⁴⁶. Są też narażeni na poważne krzywdy psychologiczne, co może wpływać na ich zdrowie psychiczne, na przykład prowadzić do depresji, stanów lękowych lub myśli o samookaleczeniu się. TFGBV często przyczynia się do wycofywania się przez jednostki z internetu, co przekłada się na niemożność pełnego korzystania z praw człowieka, w tym prawa do edukacji, do stowarzyszania się i pokojowych zgromadzeń, do udziału w życiu społecznym, kulturalnym i politycznym, zdrowia, adekwatnych standardów życiowych, pracy oraz rozwoju społecznego i gospodarczego¹⁴⁷.

Amnesty International posługuje się terminem TFGBV w odniesieniu do przemocy wobec osób LGBTI, ponieważ mowa tu o przemocy ze względu na płeć rozumianą jako zestaw konstruowanych społecznie norm, ról i zachowań powiązanych z płcią przypisaną jednostce po urodzeniu i służących utrzymywaniu cisheteropatriarchatu¹⁴⁸. Choć GBV nieproporcjonalnie częściej dotyka kobiet i dziewczynek, ma również wpływ na innych, o ile jej zasadniczą przyczyną jest dążenie do utrzymania ról, norm, systemów społecznych i systemów władzy powiązanych z płcią.

4.7 ZAKAZ NAWOŁYWANIA DO NIENAWIŚCI W MIĘDZYNARODOWYCH PRAWACH CZŁOWIEKA

Prawo międzynarodowe i standardy praw człowieka zakazują nawoływania do nienawiści, choć wypowiedzi o takim charakterze muszą być rozpatrywane zarówno w świetle prawa do równości i wolności od

¹³⁸ OHCHR, Report: *Discriminatory Laws and Practices and Acts of Violence Against Individuals Based on Their Sexual Orientation and Gender Identity*, 17 listopada 2011, UN Doc. A/HRC/19/41, akapit 20.

¹³⁹ The Yogyakarta Principles +10, zasada 36.

¹⁴⁰ UN Special Rapporteur on violence against women, Report: *Online violence against women and girls from a human rights perspective*, 18 czerwca 2018, UN Doc. A/HRC/38/47, akapit 22.

¹⁴¹ UN Special Rapporteur on violence against women, Report: *Online violence against women and girls from a human rights perspective*, akapit 67.

¹⁴² CEDAW Committee, General Recommendation 35, akapit 29.

¹⁴³ CEDAW Committee, General Recommendation 35, akapit 31 (III).

¹⁴⁴ CEDAW Committee, General Recommendation 35, akapit 12.

¹⁴⁵ CEDAW Committee, General Recommendation 35, akapit 15.

¹⁴⁶ Amnesty International, *Human Rights Implications of Technology-Facilitated Gender-Based Violence: Submission to the Human Rights Council Advisory Committee* (Index: IOR 40/9284/2025), 24 kwietnia 2025, <https://www.amnesty.org/en/documents/ior40/9284/2025/en>.

¹⁴⁷ Tamże.

¹⁴⁸ Cisheteropatriarchat to system społeczny, w którym cisplacowi, heteroseksualni mężczyźni zajmują dominujące pozycje gwarantujące im władzę i przywileje oraz możliwość wpływania na instytucje i normy społeczne w sposób, który zapewnia im korzyści, często kosztem kobiet, osób LGBTI i innych grup marginalizowanych.

dyskryminacji, jak i w świetle prawa do wolności słowa. Prawo do wolności słowa chroni wiele form wypowiedzi, nawet wypowiedzi, które mogą zostać uznane za skrajnie obraźliwe, szokujące lub wywołujące niepokój¹⁴⁹. Nie ma ono jednak charakteru absolutnego i może być ograniczane w pewnych okolicznościach, na przykład wówczas, gdy konieczne jest chronienie praw innych osób.

Prawo do równości i wolności od dyskryminacji to „ogólna podstawa ochrony praw człowieka”¹⁵⁰. W przypadku osób, których prawo do wolności od dyskryminacji jest naruszane, trzeba wdrożyć skuteczne środki naprawcze. Potwierdza to Deklaracja z Toronto dotycząca sztucznej inteligencji i uczenia maszynowego przygotowana przez organizacje społeczne i oparta na międzynarodowych standardach praw człowieka¹⁵¹:

„Firmy i podmioty prywatne projektujące i wdrażające systemy uczenia maszynowego muszą podejmować działania gwarantujące, że jednostki i grupy będą mogły liczyć na skuteczne środki naprawcze. Przykładowo może to obejmować wdrażanie przejrzystych, niezależnych, otwartych procesów pozwalających naprawić szkody wyrządzone jednostkom lub społeczeństwu, a także tworzenie w organizacjach odpowiednich stanowisk w celu szybkiego reagowania na tego rodzaju zdarzenia oraz zagwarantowania, że jednostki i grupy będą miały dostęp do skutecznych procedur odwoławczych, w tym procedur sądowych”¹⁵².

Artykuł 20. Międzynarodowego Paktu Praw Obywatelskich i Politycznych (ICCPR) precyzuje, że nawoływanie do nienawiści to nie tylko wyrażanie poglądów lub opinii wrogich wobec członków określonej grupy. To również oczywisty zamiar podżegania innych do dyskryminacji, przejawianie wrogości (silnych, irracjonalnych zachowań i emocji, takich jak potępienie, niechęć lub pogarda) lub dopuszczanie się przemocy wobec tej grupy. Przepisy zakazujące podżegania do nienawiści nie mogą naruszać prawa do wolności słowa, a ponadto muszą spełniać wymogi legalności, konieczności i proporcjonalności.

Plan działań z Rabatu o zakazie promowania nienawiści narodowej, rasowej lub religijnej, która podlega do dyskryminacji, wrogości i przemocy wymienia sześć czynników, które należy uwzględnić, kiedy ocenia się, czy wypowiedź stanowi podżeganie do nienawiści: kontekst; status lub pozycję autora wypowiedzi; intencję; treść i formę; zasięg wypowiedzi; prawdopodobieństwo, że wypowiedź doprowadzi do wyrządzenia krzywdy jednostce lub grupie¹⁵³.

Plan działań z Rabatu rozróżnia podżeganie do przemocy, wrogości lub dyskryminacji, czyli form mowy nienawiści, które muszą zostać zakazane, od wypowiedzi, które nie naruszają prawa, a mimo to budzą niepokój, bowiem naruszają zasady tolerancji i szacunku dla przekonań innych osób¹⁵⁴.

W niniejszym raporcie „nawoływanie do nienawiści” będzie rozumiane jako wypowiedzi stanowiące podżeganie do dyskryminacji, wrogości lub przemocy, które muszą być zakazane prawnie zgodnie z Artykułem 20. ICCPR. Ponadto zostanie tu omówione upowszechnianie wypowiedzi, które nie spełniają znamion nawoływania do nienawiści, a mimo to budzą niepokój, gdyż naruszają zasady tolerancji i szacunku wobec innych osób, przez co negatywnie wpływają na możliwość korzystania z prawa do wolności od dyskryminacji i prawa do równości.

Amnesty International nie próbowała oceniać, czy konkretne wypowiedzi i treści publikowane na platformie X należy uznać za nawoływanie do nienawiści. Raport ten poświęcony jest ogólnej roli, jaką X odgrywa w naruszaniu praw osób LGBTI w Polsce ze względu na nieadekwatność działań podjętych w celu ograniczenia stwarzanego przez siebie ryzyka.

¹⁴⁹ UN Human Rights Committee (HRC), General Comment 34, Article 19: Freedoms of Opinion and Expression, 12 września 2011, UN Doc. CCPR/C/GC/34, akapit 11.

¹⁵⁰ HRC, General Comment 18: Non-Discrimination, 1989, UN Doc. RI/GEN/1/Rev.9 Vol I, akapit 1.

¹⁵¹ Amnesty International, Access Now, *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems* (Index: POL 30/8447/2018), 17 maja 2018, <https://www.amnesty.org/es/documents/pol30/8447/2018/en>.

¹⁵² Amnesty International, Access Now, *The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems*, akapit 53.

¹⁵³ UN Human Rights Council, Rabat Plan of Action, 11 stycznia 2013, UN Doc. A/HRC/22/17/Ad.4, akapit 29.

¹⁵⁴ UN Human Rights Council, Rabat Plan of Action, akapit 20.

5. ROLA SERWISU X W PROPAGOWANIU PRZEMOCY ZE WZGLĘDU NA PŁEĆ WSPOMAGANEJ TECHNOLOGICZNIE W POLSCE

„W tym miejscu nienawiść mocno rezonuje”.

Aleksandra Herzyk, aktywistka.

W niniejszym rozdziale opisano, jaką rolę odegrała platforma X w rozpowszechnianiu treści anty-LGBTI, w tym treści nacechowanych nienawiścią, w latach 2019–2025.

5.1 PRZEMOC ZE WZGLĘDU NA PŁEĆ WSPOMAGANA TECHNOLOGICZNIE W SERWISIE X

Od października 2022 roku, kiedy to Elon Musk przejął firmę X, w anglojęzycznej wersji serwisu znacznie wzrosła liczba ataków na użytkowników_użytkowniczki LGBTI¹⁵⁵. Tydzień po przejęciu platformy przez Muska fundamentaliści_fundamentalistki zaczęli testować granice serwisu w kwestii wypowiedzi wymierzonych w społeczność LGBTI¹⁵⁶. Jake Shields, były zawodnik Ultimate Fighting Championship (posiadający 34 tysięcy obserwujących na X), opublikował zdjęcie pewnej drag queen z podpisem: „Tak wygląda groomer”. Następnie dodał: „Dwa miesiące temu zawiesili mnie za tego samego tweeta. Zobaczmy, czy teraz na Twitterze naprawdę panuje wolność”¹⁵⁷. Z kolei konserwatywny podkaster Matt Walsh napisał na platformie: „Zrobiliśmy olbrzymie postępy w walce z agendą trans. W ciągu zaledwie roku odrobiliśmy wieloletnie straty poniesione przez konserwatystów. Wyzwolenie Twittera nastąpiło w idealnym momencie. Możemy teraz działać jeszcze bardziej zdecydowanie”¹⁵⁸.

W badaniu przeprowadzonym w 2022 roku wśród aktywistów i organizacji LGBTI przez Amnesty International USA oraz GLAAD i Human Rights Campaign, organizacjami zajmującymi się prawami osób LGBTI, 60% respondentów oceniło, że liczba obraźliwych i nienawistnych wypowiedzi na X wzrosła od października 2022 roku¹⁵⁹. Pozostałe 40% uznało, że poziom obraźliwych i nienawistnych wypowiedzi jest taki sam jak wcześniej¹⁶⁰. Żaden z ankietowanych nie zgłosił spadku liczby obraźliwych i nienawistnych

¹⁵⁵ Amnesty International USA, *Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk*, 9 lutego 2023, <https://www.amnestyusa.org/press-releases/hateful-and-abusive-speech-towards-lgbtq-community-surging-on-twitter-under-elon-musk/>; NBC News, *Twitter is the „most dangerous platform for LGBTQ people”*, GLAAD says.

¹⁵⁶ NBC News, *A timeline of Elon Musk’s takeover of Twitter*, 17 listopada 2022, <https://www.nbcnews.com/business/business-news/twitter-elon-musk-timeline-what-happened-so-far-rcna57532>.

¹⁵⁷ NBC News, *Far-right figures appear to be testing Twitter’s boundaries for anti-LGBTQ speech*, 2 listopada 2022, <https://www.nbcnews.com/feature/nbc-out/open-season-lgbtq-people-far-right-celebrates-liberation-twitter-rcna54542>

¹⁵⁸ Tamże.

¹⁵⁹ Amnesty International USA, *Hateful and abusive speech towards LGBTQ+ community surging on Twitter under Elon Musk*.

¹⁶⁰ Tamże.

wypowiedzi¹⁶¹. Ponadto 60% respondentów stwierdziło, że nienawistne i obraźliwe wypowiedzi wpłynęły na sposób, w jaki korzystają z platformy: między innymi rzadziej publikują posty na X, dzielą się mniejszą ilością informacji na temat swojej pracy i ograniczają krąg osób, z którymi wchodzi w interakcje na platformie¹⁶². Dodatkowo 65% respondentów przyznało, że ich zdaniem na X pojawia się więcej nienawistnych i obraźliwych wypowiedzi niż na innych platformach, z których korzystają¹⁶³.

Przedstawiciele społeczności LGBTI, z którymi Amnesty International przeprowadziła wywiady w lipcu i sierpniu 2024 roku w Polsce, podzielali wiele z powyższych opinii.

Jakub Szymik, gej mieszkający w Warszawie, powiedział Amnesty International, że ogranicza swoją aktywność i widoczność w serwisie X, aby uniknąć nienawistnych ataków:

„[...] dbam o zachowanie profesjonalizmu, dlatego skupiam się na sprawach, które nie budzą kontrowersji, lub nie dzielę się opiniami w sieci pod własnym imieniem i nazwiskiem. Prowadzę anonimowe konta, na których czuję się swobodniej, ale pod własnym nazwiskiem nie piszę o pewnych sprawach. [...] sprawdzanie wiadomości i opinii w sieci sprawia, że jestem bardziej czujny... i trudno mi się zrelaksować”¹⁶⁴.

Misza, osoba niebinarna mieszkająca w Poznaniu, twierdzi, że w jej przypadku anonimowość jest kluczowa dla możliwości dalszego korzystania z X:

„Na Twitterze używam pseudonimu zamiast prawdziwego imienia. Nie wrzucam zdjęć swojej twarzy ani twarzy członków rodziny, bo się boję”¹⁶⁵.

Maja Heban, transpłciowa aktywistka z Warszawy, powiedziała Amnesty International, że jej zdaniem na X jest więcej nienawiści niż na innych platformach społecznościowych:

„W mediach społecznościowych i na Twitterze spotykam się z ogromną nienawiścią. Nie widziałam jej dotąd w takim natężeniu. Czasami dostaję negatywne komentarze na Facebooku lub Instagramie, ale to nic w porównaniu z Twitterem. Twitter to właściwie potok deadnamingu, misgenderowania, wyzywk i grożenia śmiercią”¹⁶⁶.

Respondenci podali Amnesty International przykłady treści anty-LGBTI, które widzieli na platformie:

„Widzę wiele komentarzy o tym, że ta społeczność jest chora, że niszczy wartości narodowe... używa się emotek oznaczających wymioty, kupę itp. Pojawia się wiele wpisów zawierających publiczne groźby, na przykład »będziesz następny«”¹⁶⁷.

„[Piszą], że osoby LGBTQ trafią do komór gazowych, zwracają się do o nas per śmiecie i uważają, że należy poddać nas czystkom”¹⁶⁸.

„To nienormalni ludzie, zboczeni, chorzy. Myślę, że »chore« to najczęściej używane słowo”¹⁶⁹.

„[Piszą], że ci ludzie nie są normalni, że są przeciwni polskim wartościom rodzinnym, że niszczą polskie rodziny, że nie są ludźmi, że są ideologią”¹⁷⁰.

Aleksy powiedział Amnesty International, że pod postami organizacji walczącej o prawa osób LGBTI, w której pracuje, często pojawiały się homofobiczne i transfobiczne wpisy:

„Prywatni użytkownicy publikują homofobiczne treści lub komentarze. Mam wrażenie, że pod prawie każdym tweetem na temat parad równości pojawia się wiele obrzydliwych reakcji. Zazwyczaj dbamy o to, aby nasze media społecznościowe były bezpieczną przestrzenią, dlatego staramy się ukrywać co bardziej triggerujące komentarze, ale na Twitterze, gdzie wszystko dzieje się błyskawicznie, jest to bardzo trudne”¹⁷¹.

¹⁶¹ Tamże.

¹⁶² Tamże.

¹⁶³ Tamże.

¹⁶⁴ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

¹⁶⁵ Rozmowa wideo Amnesty International z Miszą (pseudonim), 24 lipca 2024.

¹⁶⁶ Wywiad Amnesty International z Mają Heban, 30 lipca 2024.

¹⁶⁷ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

¹⁶⁸ Rozmowa wideo Amnesty International z Miszą (pseudonim), 24 lipca 2024.

¹⁶⁹ Rozmowa wideo Amnesty International z Nathanem Bryzą, 9 sierpnia 2024.

¹⁷⁰ Rozmowa wideo Amnesty International z Jolantą Prochowicz, 29 lipca 2024.

¹⁷¹ Wywiad Amnesty International z Aleksym (pseudonim), 26 lipca 2024.

5.2 ROLA OSÓB PUBLICZNYCH I POLITYKÓW POLITYCZEK W ROZPOWSZECHNIANIU TREŚCI ANTY-LGBTI W POLSCE

Na całym świecie X jest postrzegany jako platforma, na której można włączać się w dyskusje i komentować aktualne wydarzenia¹⁷². Aleksy, pracownik polskiej organizacji zajmującej się prawami osób LGBTI, w następujący sposób przedstawił Amnesty International rolę X w polskiej infosferze:

„W Polsce Twittera używają głównie politycy i dziennikarze. W porównaniu z innymi platformami społecznościowymi jest miejscem, w którym polityk albo polityczka z dużym prawdopodobieństwem odpowiedzą ci, kiedy coś do nich napiszesz. Ponadto jest szybki. Dlatego wiele decyzji ogłasza się najpierw właśnie na Twitterze. To narzędzie dobrze nadaje się do tego, aby być na bieżąco i wywierać presję [na polityków polityczki]”¹⁷³.

Aleksy dodał, że za rządów PiS społeczność LGBTI często padała ofiarą ataków:

„Spotkaliśmy się z wieloma homofobicznymi wypowiedziami niektórych polityków. Twierdzono na przykład, że społeczność LGBTI to nie ludzie, ale ideologia”¹⁷⁴.

Jakub Szymik wyjaśnia, że politycy chętnie korzystali z X w celu podsycania nastrojów anty-LGBTI:

„Myślę, że Twitter jest trybikiem w maszynie, jednym z elementów bardzo szerokiego spektrum. Przez ostatnie osiem lat w Polsce większość mowy nienawiści w przestrzeni publicznej podsycali politycy, którzy w tym celu korzystali oczywiście z Twittera, bo umożliwia on szybkie rozpowszechnianie informacji”¹⁷⁵.

Szybkie rozpowszechnianie informacji na X, napędzane algorytmami bazującymi na zaangażowaniu użytkowników_użytkowniczek, od dawna jest cechą charakterystyczną modelu biznesowego platformy: X starał się stworzyć wrażenie, że jest serwisem informacyjnym działającym w czasie rzeczywistym¹⁷⁶. W 2009 roku, czyli w samych początkach istnienia Twittera, Biz Stone, współzałożyciel serwisu, napisał, że platforma stanie się „nowym rodzajem sieci informacyjnej”¹⁷⁷.

Andrzej Duda, prezydent Polski w latach 2015–2025, jest aktywnym użytkownikiem X, w chwili pisania tego raportu (2025 roku) miał na nim 1,9 mln obserwujących¹⁷⁸. Jego duża aktywność na X może świadczyć o znaczeniu tego serwisu w polskiej sferze politycznej, a także o tym, w jak wysokim stopniu politycy zrozumieli rolę mediów społecznościowych, z X na czele, jako kluczowej przestrzeni sporów, w której kształtuje się narracje dotyczące kwestii politycznych i społecznych.

Na X trafiły treści zawierające wypowiedzi Andrzeja Dudy o wydźwięku anty-LGBTI. Na przykład 13 czerwca 2020 roku w serwisie zaczęło krążyć nagranie, na którym Duda twierdzi, że osoby LGBTI nie są ludźmi¹⁷⁹. Filmik dobrze odzwierciedla retorykę stosowaną przez PiS przed wyborami prezydenckimi w czerwcu 2020 roku, w których prawa osób LGBTI stały się jednym z głównych tematów politycznych¹⁸⁰. Oprócz krążących w sieci nagrań związanych z wyborami treści anty-LGBTI były też publikowane bezpośrednio na platformie przez samych polityków i polityczki. Na przykład Joachim Brudziński, poseł do Parlamentu Europejskiego z ramienia PiS i szef kampanii prezydenckiej Dudy, napisał na Twitterze w czerwcu 2020 roku, że „Polska bez LGBT jest najpiękniejsza”¹⁸¹. W czasie tworzenia niniejszego raportu post Brudzińskiego jest nadal dostępny na X.

Po zwycięstwie kandydata PiS w wyborach prezydenckich w 2020 roku i rozpoczęciu drugiej kadencji przez Andrzeja Dudę, ważne postaci polskiego populistycznego ruchu zwalczającego prawa LGBTI

¹⁷² X, *How many people come to Twitter for news? As it turns out, a LOT*, 12 września 2022, https://blog.x.com/en_us/topics/insights/2022/how-many-people-come-twitter-for-news.

¹⁷³ Wywiad Amnesty International z Aleksym (pseudonim), 26 lipca 2024.

¹⁷⁴ Wywiad Amnesty International z Aleksym (pseudonim), 26 lipca 2024.

¹⁷⁵ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

¹⁷⁶ The Verge, *How Twitter broke news*, 12 grudnia 2023, <https://www.theverge.com/c/features/23993135/twitter-breaking-news-history>.

¹⁷⁷ X, *What's Happening?*, 19 listopada 2009, https://blog.x.com/en_us/a/2009/whats-happening

¹⁷⁸ Profil Andrzeja Dudy na X, <https://x.com/AndrzejDuda>.

¹⁷⁹ Bartosz T. Wieliński, post na platformie X, 13 czerwca 2020, https://x.com/Bart_Wielinski/status/1271773086833094656.

¹⁸⁰ Politico, *Poland's LGBTQ community in the political crosshairs*, 19 czerwca 2020, <https://www.politico.eu/article/poland-lgbtq-community-in-the-political-crosshairs-elections-duda>.

¹⁸¹ Joachim Brudziński, post na platformie X, 11 czerwca 2020, <https://x.com/jbrudzinski/status/1271167309269413892>.

dalej publikowały na X treści wymierzone w tę społeczność; prawdopodobnie ośmielił je sukces retoryki wyborczej¹⁸². I tak na przykład dwa tygodnie po wyborach znany prawicowy komentator Rafał Ziemkiewicz (który w maju 2025 roku miał 334 200 obserwujących na X) opublikował tweet o akcji aktywistów LGBTI, którzy owinęli tęcza flagą pomnik Jezusa: „Wszystkich którzy czy to z podłości czy z naiwności wciskają ten kit że baner sodomitów narzucony na Jezusa »nie obraża« zachęcam by wrzucili wieprzowinę do meczetu a potem poszli do braci muslimów z gadką, że świnki są fajne i milusie, a zdrowe mięsko nikogo nie może obrazić”¹⁸³. W chwili tworzenia niniejszego raportu post pozostaje dostępny na X.

5.3 ANALIZA: JAK PRZEMOC ZE WZGLĘDU NA PŁEĆ WSPOMAGANA TECHNOLOGICZNIE W POLSCE PRZEKŁADA SIĘ NA REALNE KRZYWDY

Treści anti-LGBTI – podobnie jak inne treści skierowane przeciwko grupom marginalizowanym – nie funkcjonują wyłącznie w „próżni” przestrzeni internetowej, szczególnie oddzielonej od świata offline.

Jolanta Prochowicz, lesbijka mieszkająca w Lublinie, opisała to w rozmowie z Amnesty International następująco:

„Powinniśmy traktować media społecznościowe jak część naszego życia społecznego. To, co piszemy w internecie, może krzywdzić tak samo jak słowa wypowiedziane w rozmowie. To szkodliwe, bolesne, czasami bardzo. Media społecznościowe nie tyle wpływają na nasze normalne życie, ile po prostu są jego częścią. I dlatego mają na nas wpływ”¹⁸⁴.

W 2024 roku w raporcie GLAAD, organizacji zajmującej się ochroną praw osób LGBTI, zauważono: „Wielokrotnie dostrzegaliśmy istnienie bezpośredniego związku między niebezpieczną retoryką w internecie a agresywnymi zachowaniami wobec społeczności LGBTI w świecie rzeczywistym”¹⁸⁵. Podobne wnioski znalazły się w raporcie amerykańskiego Głównego Urzędu Kontroli (Government Accountability Office) z 2024 roku dotyczącym ekstremizmu w sieci: „Badania sugerują, że występowanie przestępstw z nienawiści jest powiązane z mową nienawiści w internecie, i wskazują na to, że osoby zradykalizowane online mogą stosować realną przemoc [w świecie rzeczywistym]”¹⁸⁶. Także w jednym z wcześniejszych raportów Amnesty International stwierdzono, że szkody wyrządzane przez TFGBV nie ograniczają się wyłącznie do przestrzeni cyfrowej, a ciągłość przemocy ze względu na płeć w świecie online i offline świadczy o tym, iż groźby w sieci często mogą mieć realne konsekwencje, z przemocą fizyczną włącznie¹⁸⁷.

Badacze akademicy twierdzą, że mowa nienawiści ma również poważne konsekwencje na poziomie społecznym, może bowiem „zatruci społeczeństwa, zagrażając prawom jednostki, równości i godności ludzkiej, wzmacniać napięcia między grupami społecznymi, zakłócać porządek publiczny oraz zagrażać pokojowemu współistnieniu”¹⁸⁸.

Przemoc ze względu na płeć wspomagana technologicznie (TFGBV) należy rozumieć jako zjawisko znajdujące się w tym samym spektrum co przemoc ze względu na płeć (GBV) *per se*. Istnieje w kontinuum łączącym świat online i offline, a zatem niektóre formy TFGBV mogą skutkować fizycznym naruszeniem praw

¹⁸² Global Project Against Hate and Extremism (GPAHE), *Twitter, YouTube allowing anti-LGBT hate speech, antisemitism and racism to thrive in Poland*, 18 sierpnia 2020, <https://globalextrremism.org/post/twitter-youtube-allowing-antisemitism-racism-and-anti-lgbt-hate-speech-to-thrive-in-poland/>.

¹⁸³ Rafał Ziemkiewicz, post na platformie X, 12 sierpnia 2020, https://x.com/R_A_Ziemkiewicz/status/1293565844128178177.

¹⁸⁴ Rozmowa wideo Amnesty International z Jolantą Prochowicz, 29 lipca 2024.

¹⁸⁵ GLAAD, *Social Media Safety Index*, 2024, <https://glaad.org/smsi/social-media-safety-index-2024/>, s. 2.

¹⁸⁶ United States Government Accountability Office, *Online extremism: more complete information needed about hate crimes that occur on the internet*, styczeń 2024, <https://www.gao.gov/assets/d24105553.pdf>, s. 43.

¹⁸⁷ Amnesty International, *„Being Ourselves is Too Dangerous”: Digital Violence and the Silencing of Women and LGBTI Activists in Thailand* (Index: ASA 39/7955/2024), 16 maja 2024, <https://www.amnesty.org/en/documents/asa39/7955/2024/en/>; Amnesty International, *„Everybody Here is Having Two Lives or Phones”: The Devastating Impact of Criminalization on Digital Spaces for LGBTQ People in Uganda* (Index: AFR 59/8571/2024), 23 października 2024, <https://www.amnesty.org/en/documents/afr59/8571/2024/en/>.

¹⁸⁸ Judit Bayer i in., *Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States*, 18 listopada 2021, Parlament Europejski, Komisja LIBE, Departament Tematyczny ds. Praw Obywatelskich i Spraw Konstytucyjnych, <http://dx.doi.org/10.2139/ssrn.3409279>.

w świecie rzeczywistym¹⁸⁹. Ostatnie badania przeprowadzone przez Centrum Badań nad Nienawiścią na Uniwersytecie w Leicester w Wielkiej Brytanii wykazały, że chociaż większość nastrojów anty-LGBTI w Polsce przejawia się w internecie, ich skutki są widoczne również „w realu”, w przestrzeni publicznej. Na przykład w kilku miastach w miejscach publicznych umieszczono naklejki o charakterze homofobicznym; w części wlepek ktoś ukrył żyletki z zamiarem zranienia osób próbujących je usunąć¹⁹⁰.

Co więcej, wiele odmian TFGBV, takich jak stalking i nękanie w sieci, ma swoje odpowiedniki w świecie rzeczywistym, a przemoc online – zwłaszcza gdy jest normalizowana przez osoby publiczne – może przejawiać się jako przemoc w świecie rzeczywistym¹⁹¹.

Jakub Szymik, aktywista na rzecz praw osób LGBTI, ma poczucie, że nadal musi się angażować w walkę z nienawistnymi treściami w serwisie X, mimo że ma to zły wpływ na jego samopoczucie:

„[...] uważam, że sprawdzanie wiadomości i opinii w sieci sprawia, że jestem bardziej czujny, a zarazem trudniej mi się odprężyć. Czuję, że muszę być cały czas online, bo to przygotowuje mnie na dziwne, niespodziewane sytuacje, które mogą mnie spotkać w prawdziwym życiu”¹⁹².

Aleksy, pracownik polskiej organizacji LGBTI, wyraził swoje obawy dotyczące wpływu TFGBV w serwisie X na prawo młodych ludzi do wolności wypowiedzi:

„Niektóre osoby niepełnoletnie nie są pewne lub boją się ujawnić swoją orientację seksualną: myślę, że widok tych wszystkich nienawistnych treści dodatkowo im to utrudnia”¹⁹³.

Jakub Szymik powiedział Amnesty International, że rzadko korzysta z platformy X z powodu nienawiści, którą tam widzi:

„Wydaje mi się dziś, że to głównie źródło opinii, a zwłaszcza spolaryzowanych poglądów, co zniechęca mnie do korzystania z tego serwisu”¹⁹⁴.

Przemoc ze względu na płeć wspomagana technologicznie przyczyniła się również do znacznego cierpienia psychicznego niektórych osób LGBTI w Polsce. Psycholożka Julia Kata powiedziała Amnesty International, że osoby zwracające się po wsparcie psychologiczne do walczącej o prawa LGBTI Fundacji Trans-Fuzja często wymieniają nienawiść online jako jeden z problemów, z którymi się borykają:

„Chodzi zwłaszcza o tych, dla których jedynym poczuciem wspólnoty, wsparcia i akceptacji jest internet – jeśli w sieci są wystawieni na działanie tego typu treści, po prostu tracą poczucie bezpieczeństwa. Czują się jeszcze bardziej wyobcowani: wyobcowani w prawdziwym życiu, a potem w miejscu, które uważali za swoją bezpieczną przestrzeń”¹⁹⁵.

Kwestie związane ze zdrowiem psychicznym polskiej społeczności LGBTI poruszył również Niezależny Ekspert ONZ ds. ochrony przed przemocą i dyskryminacją ze względu na orientację seksualną i tożsamość płciową. Po wizycie w Polsce w 2024 roku zauważył, że wyzwania te wiążą się:

„z utrwaloną dyskryminacją, przypadkami przemocy lub groźbami przemocy, a także z ostracyzmem społecznym i stygmatyzacją”¹⁹⁶.

Julia Kata zauważyła, że doświadczenia osób LGBTI z TFGBV potęgują ich ogólne poczucie bezbronności spowodowane dyskryminacją i przemocą:

„Ludzie nie chcą wchodzić w interakcje z innymi. Tak jak w realu. Boją się, że usłyszą coś krzywdzącego lub nienawistnego albo że ktoś będzie po prostu niemiły. To zachowania alienujące na wielu poziomach”¹⁹⁷.

¹⁸⁹ Biuro Analiz Parlamentu Europejskiego, *Combating gender-based violence: cyber violence*, marzec 2021 r., [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU\(2021\)662621_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf).

¹⁹⁰ University of Leicester, *Centre for Hate Studies submits report on Anti-LGBTI hate crime in Poland for United Nations Review*, 29 października 2024, <https://le.ac.uk/news/2024/october/lgbt>.

¹⁹¹ Biuro Analiz Parlamentu Europejskiego, *Combating gender-based violence: cyber violence*.

¹⁹² Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

¹⁹³ Wywiad Amnesty International z Aleksym (pseudonim), 26 lipca 2024.

¹⁹⁴ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

¹⁹⁵ Wywiad Amnesty International z Julią Katą, 29 lipca 2024.

¹⁹⁶ United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland, 18–29 listopada 2024, *End of mission statement*, akapit 23.

¹⁹⁷ Wywiad Amnesty International z Julią Katą, 29 lipca 2024.

Rozpowszechnianie treści anty-LGBTI odgrywa kluczową rolę w normalizacji ideologii, która dehumanizuje osoby z tej społeczności mieszkające w Polsce, a czasami też podlega do przemocy i dyskryminacji wobec nich. Masowe rozpowszechnianie tego rodzaju treści w ważnym medium społecznościowym, jakim jest X, w dużym stopniu przyczyniło się do stygmatyzacji polskiej społeczności LGBTI – stygmatyzacji, która nadal ma negatywne skutki z punktu widzenia praw osób LGBTI w tym kraju.

5.4 ZWOLNIENIA PERSONELU ODPOWIEDZIALNEGO ZA MODERACJĘ TREŚCI

Rosnąca liczba treści zawierających mowę nienawiści na platformie X może częściowo wynikać z drastycznych cięć kadrowych przeprowadzonych przez właściciela firmy X, a następnie jej prezesa, Elona Muska; jego decyzje doprowadziły do niedoboru moderatorów treści na platformie¹⁹⁸. W początkowym okresie po przejęciu firmy przez Muska, kiedy zwolniono kadrę kierowniczą i połowę pracowników, utworzono Radę ds. Moderacji Treści, której zadaniem było rozpatrywanie wniosków o przywrócenie kont¹⁹⁹.

Wkrótce po zakupie firmy Musk rozwiązał Radę ds. Zaufania i Bezpieczeństwa – grupę doradcą składającą się z prawie 100 organizacji społeczeństwa obywatelskiego, organizacji praw człowieka i innych podmiotów, które zajmowały się problemami wykorzystywania dzieci, mową nienawiści oraz treściami mogącymi prowadzić do samobójstw i samookaleczeń obecnymi na platformie²⁰⁰. Szacuje się, że zwolnił również 80% pracowników zajmujących się kwestiami zaufania i bezpieczeństwa²⁰¹. Pomimo tych cięć, gdy chodzi o obronę i poszanowanie praw osób korzystających z usług X, firma nadal uznaje znacznie osłabioną Radę za kluczowego partnera dla interesariuszy pozostającego „w ścisłej współpracy z działem produktu, działem technicznym, działami obsługi użytkowników, sprzedaży, polityki publicznej i prawnym, co ma pomóc platformie X w podejmowaniu decyzji z uwzględnieniem potrzeb osób i organizacji korzystających z [...] usług”²⁰².

Niegdyś w firmie funkcjonował także specjalny zespół zajmujący się zwalczaniem skoordynowanych kampanii dezinformacyjnych, ale eksperci i byli pracownicy twierdzą, że większość zatrudnionych w nim specjalistów zrezygnowała z pracy lub została zwolniona po październiku 2022 roku²⁰³.

Ekspert ds. praw cyfrowych wyraził obawy, że zwolnienia mogą odbić się na zdolności platformy do monitorowania szkodliwych treści, w tym różnych form nękania; z kolei reklamodawcy oświadczyli, że wycofają się z powodu własnej „niepewności” co do strategii Elona Muska²⁰⁴.

Pod koniec 2022 roku pojawiły się doniesienia, że Musk planuje w dużym stopniu polegać na zautomatyzowanej moderacji treści, zrezygnować z części audytów manualnych i postawić na ograniczenia dystrybucji bazującej na algorytmach zamiast usuwania niektórych wpisów²⁰⁵. Takie podejście, określane terminem „filtrowania widoczności”, polega na pozostawianiu niektórych tweetów naruszających zasady firmy przy jednoczesnym blokowaniu ich wyświetlania w takich miejscach jak strona główna serwisu i wyszukiwarka²⁰⁶.

¹⁹⁸ NBC News, *Twitter is the „most dangerous platform for LGBTQ people”*, GLAAD says.

¹⁹⁹ ABC News, *A timeline of Elon Musk's tumultuous Twitter acquisition*, 11 listopada 2022, <https://abcnews.go.com/Business/timeline-elon-musks-tumultuous-twitter-acquisition-attempt/story?id=86611191>.

²⁰⁰ Amnesty International, *Twitter's decision to disband safety council threatens wellbeing of users*, 13 grudnia 2022, <https://www.amnesty.org/en/latest/news/2022/12/global-twiters-decision-to-disband-safety-council-threatens-wellbeing-of-users/>.

²⁰¹ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*, 6 lutego 2024, <https://fortune.com/2024/02/06/inside-elon-musk-x-twitter-austin-content-moderation/>.

²⁰² X, *Defending and respecting the rights of people using our service*, <https://help.x.com/en/rules-and-policies/defending-and-respecting-our-users-voice>, dostęp: 3 lipca 2025.

²⁰³ BBC News, *Twitter pulls out of voluntary EU disinformation code*, 27 maja 2023, <https://www.bbc.co.uk/news/world-europe-65733969>.

²⁰⁴ ABC News, *Potential mass layoffs at Twitter could cripple content moderation, some experts say*, 23 października 2022, <https://abcnews.go.com/Business/potential-mass-layoffs-twitter-cripple-content-moderation-experts/story?id=91856973>; NBC News, *A timeline of Elon Musk's takeover of Twitter*.

²⁰⁵ Reuters, *Exclusive: Twitter leans on automation to moderate content as harmful speech surges*, 5 grudnia 2022, <https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/>.

²⁰⁶ Reuters, *Exclusive: Twitter leans on automation to moderate content as harmful speech surges*.

Na początku 2023 roku X rzekomo zaczął planować wdrożenie nowego systemu, który miałby usuwać z platformy najbardziej „niepożądane” lub szkodliwe treści²⁰⁷. Nowy plan zakładał stworzenie mniejszego, wewnętrznego zespołu moderatorów treści, z siedzibą w nowym Centrum Doskonałości ds. Zaufania i Bezpieczeństwa (Trust and Safety Center of Excellence) w Austin, w Teksasie. Miałyby on odgrywać rolę specjalistycznej sieci bezpieczeństwa zatrzymującej najbardziej rażące treści bez nadmiernego ograniczania wolności słowa, która pozostaje dla X priorytetem²⁰⁸.

Według medialnych doniesień Centrum miało zatrudniać 100 moderatorów treści, czyli znacznie mniej, niż planowano pierwotnie (500 osób), lub niż wynosiła liczba moderatorów treści zatrudnionych przez X w 2020 roku (około 1500 osób)²⁰⁹. Przed otwarciem Centrum nie było jasne, czy firma zdołała zatrudnić w nim więcej niż kilkanaście osób²¹⁰.

Także polityka X dotycząca treści szkodliwych, w tym takich, które mogą spełniać znamiona TFGBV, uległa zmianie od czasu przejścia serwisu przez Elona Muska. Na przykład w kwietniu 2023 roku serwis po cichu usunął ze swojej polityki zapisy dotyczące „umyślnego misgenderowania lub deadnamingu osób transpłciowych”²¹¹. Zapisy te przywrócono w 2024 roku²¹².

Elon Musk już wcześniej zapowiedział, że złagodzi zasady dotyczące treści niedozwolonych na platformie, sugerując, że X powinien zezwalać na wszystkie wypowiedzi, które nie naruszają prawa krajowego państw, w których działa²¹³.

W rozmowie z magazynem Fortune osoba zorientowana w kwestiach zaufania i bezpieczeństwa w serwisie X zwróciła uwagę na wpływ niejasnych lub szybko zmieniających się zasad w firmie na zdolność platformy do odpowiedniego reagowania na szkodliwe treści: „W pewnym sensie ważniejsze niż liczba osób pracujących przy komputerach są przejrzyste zasady oparte na strategii ograniczania zachowań szkodliwych oraz narzędzia i systemy niezbędne do wdrażania tych zasad na dużą skalę”²¹⁴.

Krytycy twierdzą ponadto, że polityka X w zakresie dezinformacji i technologie wykrywania pozwoliły na promowanie na platformie postów podżegających do nienawiści. Częściowo dzieje się tak dlatego, że tak zwane konta Twitter Blue należące do użytkowników_użytkowniczek, którzy zapłacili za weryfikację, są obecnie promowane przez algorytm serwisu, nawet jeśli są wykorzystywane do rozpowszechniania fałszywych lub szkodliwych treści²¹⁵.

Niezależny audyt oceny ryzyka przeprowadzony na mocy Aktu o usługach cyfrowych (DSA) wykazał brak konsekwencji w stosowaniu zasad moderacji treści²¹⁶. Wydaje się, że konta z dużą liczbą obserwujących lub statusem „zweryfikowane” traktuje się inaczej niż konta zwykłych użytkowników_użytkowniczek²¹⁷. W audycie zalecono, aby X wdrożył standardowe procedury egzekwowania przepisów w celu zapewnienia wszystkim użytkownikom_użytkowniczkom sprawiedliwego i transparentnego traktowania²¹⁸.

²⁰⁷ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*.

²⁰⁸ Tamże.

²⁰⁹ Paul M. Barrett, *Who moderates the social media giants? A call to end outsourcing*, czerwiec 2020, <https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/nyu-content-moderation-report-june-8-2020.pdf>.

²¹⁰ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*.

²¹¹ NBC News, *Twitter is the „most dangerous platform for LGBTQ people”, GLAAD says*.

²¹² GLAAD, *Social Media Safety Index*.

²¹³ ABC News, *Potential mass layoffs at Twitter could cripple content moderation, some experts say*; NBC News, *Elon Musk, new owner of Twitter, tweets unfounded anti-LGBTQ conspiracy theory about Paul Pelosi attack*, 30 października 2022, <https://www.nbcnews.com/news/us-news/elon-musk-new-owner-twitter-tweets-unfounded-conspiracy-theory-paul-pe-rcna54717>.

²¹⁴ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*.

²¹⁵ Foreign Policy, *Elon Musk's Twitter is becoming a sewer of disinformation*, 15 lipca 2023, <https://foreignpolicy.com/2023/07/15/elon-musk-twitter-blue-checks-verification-disinformation-propaganda-russia-china-trust-safety/>; *The Independent*, *How Twitter became a far-right misinformation vehicle for the riots – and how it could be saved*, 6 sierpnia 2024, <https://www.independent.co.uk/tech/twitter-x-riots-elon-musk-b2592074.html>.

²¹⁶ FTI Consulting, *„X Independent Audit”: Article 37 of the Regulation (EU) 2022/2065 (DSA)*, 2024, <https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-audit/TIUC-DSA-Audit-Report-2024-08-27.pdf>.

²¹⁷ Tamże.

²¹⁸ Tamże.

5.4.1 UWAGI SPOŁECZNOŚCI (COMMUNITY NOTES)

Od 2021 roku X w coraz większym stopniu polega na funkcji „Uwagi społeczności”, która umożliwia zdecentralizowanej siatce zweryfikowanych użytkowników_użytkowniczek umieszczanie pod wpisami notatek z dodatkowym kontekstem: jest to forma moderacji treści dla milionów aktywnych użytkowników_użytkowniczek²¹⁹. Funkcja „Uwagi społeczności” ma ponad 100 tysięcy autorów_autorek w całej UE; w Polsce została udostępniona w 2023 roku²²⁰.

Autorzy_autorki uwag społeczności rozpoczynają od oceniania uwag innych użytkowników, po czym uzyskują możliwość samodzielnego pisania własnych notatek kontekstowych²²¹. Autorem_autorką może zostać każdy posiadacz_każda posiadaczka konta X cieszącego się „dobrą reputacją”, którzy nie naruszają zasad X i są właścicielami zweryfikowanego numeru telefonu. Autorzy_autorki uwag są wybierani przez X co tydzień w sposób losowy²²².

Autorzy_autorki oceniają przydatność uwag na podstawie szeregu kryteriów heurystycznych, takich jak neutralność języka, zawartość wysokiej jakości cytatów lub bezpośrednie odniesienie do treści wpisu²²³.

Decyzja o umieszczeniu uwagi pod postem należy wyłącznie do zweryfikowanych użytkowników_użytkowniczek. Pracownicy X nie umieszczają ani nie usuwają uwag²²⁴. Od lipca 2024 roku użytkownicy_użytkowniczki mogą poprosić o dodanie uwag społeczności do wpisu, który ich zdaniem wymaga dodatkowego kontekstu²²⁵. Każdy post i każdy użytkownik_użytkowniczka mogą zostać objęci procedurą uwag społeczności; dotyczy to także reklamodawców. Treści, które ktoś opatruje uwagą, zostają zdemonetyzowane, co oznacza, że autor_autorka nie będzie mógł_mogła uzyskać przychodów z danego wpisu²²⁶.

Autorzy_autorki postów mogą wystąpić o weryfikację uwagi dołączonej do ich treści; dokonują jej autorzy_autorki uwag społeczności²²⁷. Jeśli ponowna weryfikacja zmieni ocenę uwagi społeczności na „nieprzydatną”, zostanie ona usunięta²²⁸. X nie nadzoruje procesu weryfikacji²²⁹. Jeśli uwaga jest widoczna przez dwa tygodnie, zostaje na stałe przypisana do postu i już nie można jej usunąć²³⁰.

Skuteczność opisanych powyżej metod przeciwdziałania szkodliwym treściom na platformie została skrytykowana przez organizacje zajmujące się ochroną praw cyfrowych. Pojawiły się również obawy dotyczące jakości niektórych zgłoszeń oraz tego, czy takie podejście nadąży za tempem tworzenia treści na X²³¹. Firma twierdzi, że według niezależnych badań konta lub wpisy, które opatrzone uwagami społeczności, są o 80% bardziej narażone na usunięcie, a prawdopodobieństwo ich udostępnienia jest od 50% do 61% mniejsze²³².

²¹⁹ Tech Crunch, *Elon Musk takes Twitter out of the EU's Disinformation Code of Practice*, 27 maja 2023, <https://techcrunch.com/2023/05/27/elon-musk-twitter-eu-disinformation-code/>; Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*; House of Commons Science Innovation and Technology Committee, *Oral evidence: social media, misinformation and harmful algorithms*, HC 441, 25 lutego 2025, <https://committees.parliament.uk/oralevidence/15413/pdf/>.

²²⁰ Community Notes, post na platformie X, 27 lipca 2023, <https://x.com/communitynotes/status/1684667112869052416>; X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024, <https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-sra/dsa-sra-2024/TIUC-DSA-SRA-Report-2024.pdf>.

²²¹ House of Commons Science Innovation and Technology Committee, *„Oral evidence: Social media, misinformation and harmful algorithms”*, HC 441.

²²² Tamże.

²²³ Tamże.

²²⁴ Tamże.

²²⁵ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

²²⁶ House of Commons Science Innovation and Technology Committee, *Oral evidence: Social media, misinformation and harmful algorithms*, HC 441.

²²⁷ X, Community Notes, <https://communitynotes.x.com/guide/en/contributing/additional-review>.

²²⁸ X, Community Notes.

²²⁹ X, Community Notes.

²³⁰ House of Commons Science Innovation and Technology Committee, *Oral evidence: Social media, misinformation and harmful algorithms*, HC 441.

²³¹ Tech Crunch, *Elon Musk takes Twitter out of the EU's Disinformation Code of Practice*.

²³² House of Commons Science Innovation and Technology Committee, *Oral evidence: Social media, misinformation and harmful algorithms*, HC 441.

Chociaż zdaniem X zmniejsza się dzięki temu ryzyko, że takie wpisy staną się wiralami, firma oznajmiła również, że wpływa na zasięgi postów opatrzonych uwagami społeczności²³³. Oznacza to, że rzadsze udostępnianie postów z uwagami społeczności wynika z zachowania użytkowników_użytkowniczek, a nie algorytmicznego obniżania ich statusu (tzw. downrankingu).

Ponadto funkcja „Uwagi społeczności” działa wolno. Uwaga jest umieszczana pod postem tylko wtedy, gdy dwoje autorów_auterek, którzy wcześniej nie zgadzali się ze sobą, uzna, że w danym przypadku taki komentarz będzie przydatny. Serwis X przyznał, że stwarza to problemy związane z szybkością i skalą²³⁴. Chociaż firma informuje, że szybkość działania funkcji CN rośnie, średni czas pojawienia się uwagi po opublikowaniu wpisu w sytuacji kryzysowej – np. ataku Hamasu na Izrael 7 października 2023 roku – wynosi około pięciu godzin, co potencjalnie daje setkom tysięcy osób czas na przeczytanie postu przed dodaniem uwagi²³⁵. Badanie przeprowadzone w 2021 roku na podstawie danych serwisu Birdwatch, poprzednika Community Notes, wykazało, że użytkownicy_użytkowniczki są bardziej skłonni do negatywnego oceniania tweetów osób, z którymi nie zgadzają się w kwestiach politycznych, oraz do oceniania wpisów osób o odmiennych poglądach politycznych jako nieprzydatnych – co pozwala przypuszczać, że użytkownicy_użytkowniczki chętniej kwestionują treści osób, z którymi nie zgadzają się pod względem ideologicznym²³⁶. Chociaż taki stan rzeczy niekoniecznie oznacza, że przyjęta metoda weryfikacji faktów nieskuteczne identyfikuje wpisy wprowadzające w błąd, to sugeruje jednak, że upolitycznienie może odgrywać znaczącą rolę w ocenie treści²³⁷. W tym samym badaniu z 2021 roku stwierdzono też, że użytkownicy_użytkowniczki, którzy obserwowali konta o podobnych poglądach politycznych do autora_autorki posta – a zatem mogli częściej natknąć się na dany post w swoim kanale informacyjnym – byli bardziej skłonni do krytycznego oznaczania uwag (tj. oceniania ich jako nieprzydatne), które informują o tym, że rzeczony wpis wprowadza czytelników w błąd²³⁸. W związku z tym osoby, co do których istnieje duże prawdopodobieństwo, że wyświetliły uwagę społeczności dotyczącą danego tweeta, były najmniej skłonne uznać taką uwagę za przydatną²³⁹.

Uwagi społeczności, w połączeniu z drastycznym obcięciem środków przeznaczanych na moderację, wydają się próbą uchylenia się od odpowiedzialności za kwestie związane z prawami człowieka: w zasadzie X „podzleca” moderowanie treści umieszczanych w serwisie własnym użytkownikom_użytkowniczkom. Jest to działanie ewidentnie niezgodne z normami praw człowieka: brak jasnych zasad i polityki dotyczących treści opatrywanych uwagami społeczności oznacza, że użytkownicy_użytkowniczki nie są w stanie dokładnie przewidzieć, czy ich treści zostaną oznaczone.

Ponadto wydaje się, że funkcja „Uwagi społeczności” nie uwzględnia w wystarczającym stopniu kwestii legalności, konieczności i zasadności, które trzeba brać pod uwagę przy ograniczaniu wolności wypowiedzi zgodnie z międzynarodowymi normami praw człowieka, takimi jak artykuł 19. Międzynarodowego paktu praw obywatelskich i politycznych oraz wytyczne inicjatywy GNI. Dzieje się tak, ponieważ do oznaczenia wpisu wystarczy zgoda dwóch osób, które wcześniej miały odmienne zdanie. Co więcej, funkcja „Uwagi społeczności” nie zapewnia żadnej dodatkowej ochrony praw grup marginalizowanych.

Funkcja Community Notes pozwala w praktyce na moderowanie treści na platformie w oparciu o opinie użytkowników_użytkowniczek, a nie o rzetelne, jasne i starannie przemyślane zasady opracowane zgodnie z normami praw człowieka. Ponadto autorzy_autorki Community Notes nie przechodzą żadnego szkolenia w zakresie weryfikacji faktów, a treść uwag nie podlega jakiegokolwiek kontroli jakości ze strony X. Dlatego fakt, że funkcja „Uwagi społeczności” została przedstawiona jako główny środek ograniczający ryzyko w najnowszej ocenie ryzyka serwisu X przeprowadzonej na mocy DSA, budzi poważne obawy i zastrzeżenia²⁴⁰.

²³³ Tamże.

²³⁴ Tamże.

²³⁵ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

²³⁶ Jennifer Allan i in., *Birds of a feather don't fact-check each other: partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program*, CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Article No. 245, <https://dl.acm.org/doi/pdf/10.1145/3491102.3502040>.

²³⁷ Tamże.

²³⁸ Tamże.

²³⁹ Tamże.

²⁴⁰ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*.

5.5 ZMIANY PODEJŚCIA SERWISU X DO TREŚCI NIENAWISTNYCH

Wkrótce po przejściu przez Elona Muska serwisu X przywrócono konta, które zostały wcześniej zablokowane za naruszenie zasad platformy. Należały do nich na przykład konto amerykańskiej kongresmenki Marjorie Taylor Greene (zablokowane za naruszenie wytycznych dotyczących dezinformacji na temat COVID-19), konto satyrycznego serwisu informacyjnego The Babylon Bee (zablokowanego za opublikowanie transfobicznego artykułu, który naruszał zasady dotyczące zachowań nienawistnych), oraz konto Andrew Tate'a, „gwiazdy” mediów społecznościowych (zablokowanego za stwierdzenie, że kobiety powinny ponosić część odpowiedzialności za to, że padają ofiarami napaści seksualnych)²⁴¹.

Organizacje społeczeństwa obywatelskiego zajmujące się prawami osób LGBTI, takie jak amerykańska GLAAD, również dostrzegły zmianę w sposobie reagowania serwisu X na wpisy naruszające zasady platformy²⁴². Wcześniej obraźliwe posty usuwano, jednak w chwili sporządzania niniejszego raportu niektóre wpisy naruszające regulamin X podlegają jedynie ograniczeniu widoczności, nie zostają natomiast całkowicie usunięte z platformy²⁴³.

W październiku 2023 roku X wprowadził kilka zmian w wytycznych dla społeczności; obejmowały one znaczne złagodzenie zasad dotyczących treści zawierających przemoc²⁴⁴, definiowanych jako treści zawierające „język przemocy” lub „media zawierające przemoc”²⁴⁵. Język przemocy został zdefiniowany w regulaminie jako treści, które grożą, podlegają, gloryfikują lub wyrażają pragnienie przemocy lub wyrządzenia krzywdy, natomiast media zawierające przemoc to materiały wizualne przedstawiające treści drastyczne, pełne przemocy lub okrucieństwa, w tym przemoc seksualnej²⁴⁶.

Przed 30 października 2023 roku wytyczne społeczności X stanowiły: „Prowadzimy **politykę zero tolerancji**, jeśli chodzi o mowę nienawiści, żeby zapewnić bezpieczeństwo naszym użytkownikom_użytkowniczkom i zapobiec normalizowaniu przemocy”²⁴⁷. Obecnie zapis ten brzmi: „**Możemy usuwać niektóre treści lub ograniczać ich widoczność**, żeby zapewnić bezpieczeństwo naszym użytkownikom i zapobiec normalizowaniu przemocy”²⁴⁸. Ponadto w wytycznych dotyczących treści zawierających przemoc brak jakiegokolwiek wzmianki o zagrożeniach, jakie treści te stwarzają dla społeczności marginalizowanych, w tym zagrożeniach związanych z TFGVB²⁴⁹. Jednocześnie polityka X dotycząca zachowań nienawistnych zabrania atakowania osób ze względu na ich rasę, pochodzenie etniczne, narodowość, kastę, orientację seksualną, płeć, tożsamość płciową, przynależność religijną, wiek, niepełnosprawność lub poważną chorobę²⁵⁰.

W 2023 roku badacze z Centrum Przeciwdziałania Nienawiści Cyfrowej (Center for Countering Digital Hate) zgłosili do X 300 postów zawierających mowę nienawiści. Tydzień później serwis nadal udostępniał 86% tych postów²⁵¹. Badacze otrzymali od serwisu X powiadomienia, że trzy konta, które opublikowały treści nacechowane mową nienawiści, zostały zablokowane i że „użytkownicy tych kont nie mają możliwości publikowania, udostępniania ani lajkowania żadnych treści, a jeśli zechcą odzyskać pełny dostęp do swoich kont, zostaną poproszeni o usunięcie zgłoszonych wpisów”. Mimo to zgłoszone posty pozostały widoczne na platformie²⁵².

Zgodnie z zapisami DSA bardzo duże platformy internetowe (VLOP), takie jak X, muszą opracowywać raporty dotyczące transparentności, w których opisują swoje działania związane z moderowaniem treści.

²⁴¹ Fortune, *Inside the shifting plan at Elon Musk's X to build a new team and police a platform „so toxic it's unrecognisable”*.

²⁴² NBC News, *Twitter is the „most dangerous platform for LGBTQ people”, GLAAD says*.

²⁴³ Tamże.

²⁴⁴ Lab Platform Governance, Media and Technology (PGMT), *X (formerly Twitter), softens its violent speech policy*, 2 listopada 2023, <https://platform-governance.org/2023/x-formerly-twitter-softens-its-violent-speech-policy/>.

²⁴⁵ X, *Violent Content policy*, luty 2025, <https://help.x.com/en/rules-and-policies/violent-content>.

²⁴⁶ Tamże.

²⁴⁷ PGMT, *X (formerly Twitter) softens its violent speech policy* (podkreślenie Amnesty International).

²⁴⁸ X, *Violent Content policy* (podkreślenie Amnesty International).

²⁴⁹ Tamże; The Verge, *Twitter rewrites its rules on violent content under Elon Musk*, 28 lutego 2023, <https://www.theverge.com/2023/2/28/23619262/twitter-violent-speech-policy-zero-tolerance>.

²⁵⁰ X, *Hateful conduct policy*, kwiecień 2023, <https://help.x.com/en/rules-and-policies/hateful-conduct-policy>.

²⁵¹ Center for Countering Digital Hate, *X content moderation failure: how Twitter/X continues to host posts reported for extreme hate speech*, wrzesień 2023, https://counterhate.com/wp-content/uploads/2023/09/230907-X-Content-Moderation-Report_final_CCDH.pdf.

²⁵² Tamże, s. 6.

W momencie sporządzania niniejszego sprawozdania X był platformą, która wykazywała największą niespójność w swoich raportach – takie wnioski płyną z wyników akademickiego audytu Bazy Danych Transparentności DSA, która zawiera wszystkie raporty złożone przez osiem największych platform mediów społecznościowych w UE w ciągu pierwszych 100 dni funkcjonowania bazy²⁵³.

W najnowszym raporcie dotyczącym transparentności zgodnym z wytycznymi DSA firma X opisuje swoją politykę moderacji treści następująco: „została opracowana i dostosowana w celu ograniczenia systematycznego ryzyka bez zbędnego ograniczania dostępu do naszych usług i praw podstawowych, w szczególności wolności słowa. Działania związane z moderacją treści są zakorzenione w spójnych zasadach i na podstawie tychże wdrażane, obejmują także różnorodny zestaw środków służących do interwencji. W ten sposób dbamy o to, aby nasze działania były racjonalne, proporcjonalne i skuteczne. Nasze systemy moderacji treści łączą automatyczną i ludzką weryfikację z rozbudowanym systemem odwoławczym, który umożliwia użytkownikom_użytkowniczkom szybkie zgłaszanie potencjalnych nieprawidłowości lub błędów w moderacji”²⁵⁴.

Platforma przyznaje również, że „przemoc, nękanie i podobne zachowania zniechęcają ludzi do wyrażania siebie, a tym samym zmniejszają jakość globalnej debaty publicznej”²⁵⁵.

X przedstawił ponadto czynniki wpływające na decyzje dotyczące egzekwowania przepisów na platformie, do których należą następujące pytania: czy zachowanie jest skierowane przeciwko osobie, grupie lub kategorii osób chronionych? Czy zgłoszenie zostało złożone przez osobę, której dotyczy nadużycie, czy też przez osobę postronną? Czy zgłoszony użytkownik_użytkowniczka w przeszłości publikował treści naruszające przepisy, jaki był stopień tego naruszenia oraz czy temat zgłoszonego wpisu mieści się w interesie publicznym?²⁵⁶

Zgodnie z wymogami DSA raporty dotyczące transparentności bardzo dużych platform internetowych muszą zawierać dane dotyczące liczebności kadry skierowanej do pracy przy moderowaniu z uwzględnieniem tego, którym z języków urzędowych UE posługują się pracownicy. W pierwszej partii raportów, opublikowanej w listopadzie 2023 roku, ujawniono, że X zatrudniał tylko jednego moderatora treści mówiącego po polsku, a jedynie 8% moderatorów treści platformy biegle posługiwało się językiem urzędowym UE innym niż angielski²⁵⁷. W najnowszym raporcie przygotowanym na mocy DSA (w okresie sporządzania niniejszego dokumentu, który zaczął powstawać w kwietniu 2025 roku) czytamy, że X zatrudniał tylko dwóch polskojęzycznych moderatorów treści, przy czym dla jednego z nich polski był drugim językiem²⁵⁸.

X ujawnił w swoim raporcie dotyczącym przejrzystości, że gdy pojawia się potrzeba dodatkowego wsparcia językowego, firma korzysta z narzędzi tłumaczenia maszynowego²⁵⁹.

Zdaniem Jakuba Szymika brak odpowiedniej liczby moderatorów posługujących się językiem polskim spowodował opóźnienia w reagowaniu na zgłaszane na platformie szkodliwe treści:

„Zrobiłem to [zgłosiłem treść] kilka razy, ale dopiero gdy zobaczyłem treści dotyczące osoby, którą znałem osobiście. I nic nie usunęło. Dostałem powiadomienia, że treść pozostanie na stronie. Nie ma sensu tracić czasu na takie rzeczy. Na 8 milionów użytkowników w Polsce jest tylko jeden moderator polskojęzyczny... Nie mam zaufania do tej platformy... Nie zgłaszam już takich przypadków”²⁶⁰.

Również Mateusz Kaczmarek z Grupy Stonewall nie zauważył żadnych skutków swojego zgłoszenia:

„Czasem gdy korzystam z Twittera, zgłaszam niektóre tweety. Myślę, że większość tych zgłoszeń niczego nie zmienia: tamte wpisy nadal wiszą, nic się z nimi nie dzieje”²⁶¹.

Także Maja Heban powiedziała Amnesty International, że nie zgłasza już treści do X z powodu niewystarczających reakcji platformy:

²⁵³ Amaury Trujillo i in., *The DSA Transparency Database: auditing self-reported moderation actions by social media*, 1 August 2024, Proceedings of the ACM on Human-Computer Interaction, Volume 9, Issue 2, <https://dl.acm.org/doi/10.1145/3711085>.

²⁵⁴ X, *DSA Transparency Report – October 2024*, <https://transparency.x.com/dsa-transparency-report.html>.

²⁵⁵ Tamże.

²⁵⁶ Tamże.

²⁵⁷ Global Witness, *How big tech platforms are neglecting their non-English language users*, 30 listopada 2023, <https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/>.

²⁵⁸ X, *DSA Transparency Report – April 2025*, <https://transparency.x.com/dsa-transparency-report-2025-april.html>.

²⁵⁹ Tamże.

²⁶⁰ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

²⁶¹ Wywiad Amnesty International z Mateuszem Kaczmakiem, 28 lipca 2024.

„Przestałam je zgłaszać rok temu. Próbowalam na samym początku, kiedy Musk przejął serwis. Ale wszyscy szybko zdali sobie sprawę, że to już nie działa. Chociaż nawet przed Muskiem to nie była skuteczna metoda. Czasami zgłaszałam mocno wiralowe treści, ale według Twittera były w porządku. Najobrzydliwsze rzeczy, jakie można sobie wyobrazić, były w porządku. Szybko stało się jasne, że usunięcie może grozić najwyżej postom zawierający CSAM [materiały przedstawiające wykorzystywanie seksualne dzieci], ale poza takimi materiałami można napisać dosłownie wszystko. I nikt nie wyrzuci tego ze strony”²⁶².

Psycholożka Julia Kata w rozmowie z Amnesty International zauważyła, że trudno było zgłaszać również komentarze pod postami:

„Dostaje się tam mnóstwo komentarzy. Czasami łatwiej jest po prostu usunąć treść, zamiast ją akceptować. Komentarze bardzo trudno zgłosić. Serwisu to po prostu nie obchodzi”²⁶³.

Piotr Pjotrowicz, gej mieszkający w Sosnowcu, używający X w celach aktywistycznych, powiedział Amnesty, że jego zdaniem nieodpowiednia moderacja w tym serwisie wpisuje się w szerszy trend zaniedbywania potrzeb użytkowników_użytkowniczek nieanglojęzycznych przez platformy technologiczne:

„Myślę, że jest to problem globalny, a nie tylko polski. W wielu innych krajach też się tym nie przejmują i nie mają wystarczającej liczby osób, aby opanować aktualną sytuację”²⁶⁴.

Chociaż firma X twierdzi, że każdy moderator treści przechodzi kompleksowe szkolenia i kursy uzupełniające, nie jest jasne, czy obejmują one problematykę praw osób LGBTI. Spośród wszystkich firm ocenionych zgodnie z indeksem GLAAD dotyczącym bezpieczeństwa mediów społecznościowych X jako jedyna platforma nie ujawniła żadnych informacji na temat tego, czy prowadzi szkolenia dotyczące potrzeb użytkowników_użytkowniczek LGBTI dla moderatorów treści²⁶⁵. 22 sierpnia 2024 roku Amnesty International zwróciła się do X z pytaniem o typ i poziom szkoleń dotyczących praw osób LGBTI dostępnych dla moderatorów treści serwisu, jednak nie otrzymała odpowiedzi²⁶⁶.

5.6 OPÓR PRZED DOSTOSOWANIEM SIĘ DO PRZEPISÓW I STANDARDÓW UE

W 2018 roku poprzednie kierownictwo X (wówczas działającego pod nazwą Twitter) podpisało dobrowolny Europejski kodeks postępowania w zakresie dezinformacji²⁶⁷. Firma zobowiązała się tym samym do podjęcia działań mających na celu zwalczanie rozpowszechniania fałszywych informacji w ramach swoich usług poprzez ograniczenie przychodów z reklam powiązanych z takimi informacjami, zwalczanie botów i fałszywych kont, zapewnienie konsumentom narzędzi do zgłaszania dezinformacji oraz umożliwienie badaczom prowadzenia analiz dotyczących platformy i jej działania²⁶⁸.

W czerwcu 2022 roku Komisja Europejska przedstawiła ulepszoną wersję wspomnianego kodeksu i ogłosiła utworzenie Centrum Transparentności mające monitorować przestrzeganie jego postanowień²⁶⁹. Komisja Europejska ogłosiła również, że przestrzeganie kodeksu będzie jednym z wymogów zgodności przedsiębiorstw z DSA²⁷⁰. W maju 2023 roku X wycofał się z dobrowolnego przestrzegania zapisów kodeksu, wchodząc w spór z Komisją Europejską, która uznała to za „konfrontację”²⁷¹. W tym samym czasie Elon Musk utrzymywał, że od momentu przejścia przez niego X na platformie jest mniej dezinformacji niż przedtem²⁷².

²⁶² Wywiad Amnesty International z Mają Heban, 20 lipca 2024.

²⁶³ Wywiad Amnesty International z Julią Katą, 29 lipca 2024.

²⁶⁴ Rozmowa wideo Amnesty International z Piotrem Pjotrowiczem, 7 sierpnia 2024.

²⁶⁵ GLAAD, *Social Media Safety Index*; X, *DSA Transparency Report – October 2024*.

²⁶⁶ Pismo Amnesty International skierowane do firmy X, 22 sierpnia 2024.

²⁶⁷ Tech Crunch, *Elon Musk takes Twitter out of the EU's Disinformation Code of Practice*.

²⁶⁸ Tamże.

²⁶⁹ Tamże.

²⁷⁰ Tamże.

²⁷¹ BBC News, *Twitter pulls out of voluntary EU disinformation code*; Tech Crunch, *Elon Musk takes Twitter out of the EU's Disinformation Code of Practice*; Euronews, *Twitter has chosen „confrontation” with Brussels over disinformation code of conduct*, 5 czerwca 2023, <https://www.euronews.com/my-europe/2023/06/05/twitter-has-chosen-confrontation-with-brussels-over-disinformation-code-of-conduct>.

²⁷² BBC News, *Twitter pulls out of voluntary EU disinformation code*.

X został już wcześniej objęty pierwszym w historii dochodzeniem prowadzonym na podstawie przepisów DSA²⁷³. W grudniu 2023 roku Komisja Europejska wszczęła postępowanie w sprawie uchybień X po otrzymaniu wielokrotnych skarg, że platforma nie podejmuje wystarczających działań w celu ograniczenia rozprzestrzeniania dezinformacji i mowy nienawiści w internecie²⁷⁴. Wszczęto cztery dochodzenia w sprawie nieprzestrzegania przez X unijnych przepisów dotyczących przeciwdziałania nielegalnym treściom i dezinformacji, a także przepisów dotyczących przejrzystości w zakresie reklam i dostępu badaczy do danych²⁷⁵. W lipcu 2024 roku Komisja Europejska przekazała X wstępną opinię, że platforma narusza przepisy DSA w obszarach związanych z tak zwanymi ciemnymi praktykami (nieuczciwymi technikami stosowanymi przez platformy internetowe w celu manipulowania zachowaniami użytkowników_ użytkowników, często bez ich wiedzy lub zgody), transparentnością reklam oraz dostępem badaczy do danych²⁷⁶. Podobno Elon Musk często określał DSA mianem „narzędzia cenzury”²⁷⁷.

W kwietniu 2025 roku Komisja Europejska podała do wiadomości, że rozważa nałożenie na X wysokiej grzywny za naruszenie przepisów DSA. Ponadto zażądała od X zmian w sposobie funkcjonowania²⁷⁸. Nadal istnieje możliwość, że UE i X osiągną porozumienie, o ile firma zgodzi się na zmiany, które rozwieją obawy organu regulacyjnego²⁷⁹.

Firma X zareagowała na doniesienia mediów o zbliżającej się grzywnie, publikując oświadczenie, w którym określiła potencjalne działania egzekucyjne wobec niej „bezprecedensowym aktem cenzury politycznej i atakiem na wolność słowa”²⁸⁰.

Dorota Głowacka, ekspertka ds. rzecznictwa i sporów sądowych z Fundacji Panoptykon, polskiej organizacji pozarządowej zajmującej się prawami cyfrowymi, liczy na skuteczność tej drogi wyegzekwowania od X zgodności z DSA:

„Uważamy, że to jedyny skuteczny sposób, bo inne środki – samoregulację, kodeksy postępowania, wszelkiego rodzaju rozwiązania – już wypróbowaliśmy. Okazały się po prostu nieskuteczne, nie przyniosły żadnych istotnych zmian”²⁸¹.

5.7 PROBLEMY SYSTEMOWE

Chociaż od czasu przejęcia firmy przez Elona Muska w 2022 roku X podlega wzmożonej obserwacji organów Unii Europejskiej, należy zauważyć, że nawet przed zmianą właściciela firma X (wówczas znana jako Twitter) była często krytykowana za niewłaściwe monitorowanie treści potencjalnie szkodliwych, w tym treści, które spełniają znamiona TFGBV²⁸².

W 2018 roku Amnesty International stwierdziła, że X nie przestrzega praw kobiet w internecie, gdyż niewłaściwie reaguje na przypadki nadużyć wobec nich online. Odnotowano, że najbardziej narażone na przemoc na platformie były kobiety niebiałe, kobiety należące do mniejszości etnicznych lub wyznaniowych, lesbijki, kobiety biseksualne i transpłciowe, osoby niebinarne oraz kobiety niepełnosprawne²⁸³. Badania wykazały, że nawet w 2018 roku liczbę szkodliwych treści obecnych w serwisie X postrzegano jako większą niż na innych platformach.

²⁷³ Politico, *X vs. EU: Elon Musk hit with probe over spread of toxic content*, 18 grudnia 2023, <https://www.politico.eu/article/elon-musks-x-probed-by-eu-for-failing-to-curb-toxic-content/>.

²⁷⁴ Tamże.

²⁷⁵ Tamże.

²⁷⁶ Komisja Europejska, *Commission send preliminary findings to X for breach of Digital Services Act*, 12 lipca 2024, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761.

²⁷⁷ Tech Policy Press, *Understanding the EU's Digital Services Act Enforcement Against X*, 5 kwietnia 2025, <https://www.techpolicy.press/understanding-the-eus-digital-services-act-enforcement-against-x/>.

²⁷⁸ New York Times, *E.U. prepares major penalties against Elon Musk's X*, 3 kwietnia 2025, <https://www.nytimes.com/2025/04/03/technology/eu-penalties-x-elon-musk.html>.

²⁷⁹ Tamże.

²⁸⁰ Global Government Affairs, post na platformie X, 4 kwietnia 2025, <https://x.com/GlobalAffairs/status/1907963263419297893>.

²⁸¹ Wywiad Amnesty International z Dorotą Głowacką, 29 lipca 2024.

²⁸² The Independent, *How Twitter became a far-right misinformation for the riots – and how it could be saved*.

²⁸³ Amnesty International, *Toxic Twitter – a toxic place for women*, 21 marca 2018, <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/>.

Jessica Valenti, amerykańska dziennikarka i pisarka, powiedziała w wywiadzie przeprowadzonym w ramach projektu Amnesty International: „[szkodliwe] Treści na wszystkich platformach wydają się dość podobne, ale tym, co wyróżnia Twittera, jest ich ogromne natężenie”²⁸⁴.

W tym samym 2018 roku Amnesty International przeprowadziła badania dotyczące platformy X, aby poznać skalę publikowania na niej treści obraźliwych, skierowanych przeciwko dziennikarkom i kobietom zajmującym się polityką²⁸⁵. W momencie przeprowadzania badań jeden na 10 tweetów przeanalizowanych przez Amnesty International dotyczących czarnych polityczek i dziennikarek z Wielkiej Brytanii i Stanów Zjednoczonych zawierał treści obraźliwe lub budzące zastrzeżenia²⁸⁶. Wyniki te skłoniły Amnesty International do scharakteryzowania serwisu X jako „przestrzeni, w której rasizm, mizoginia i homofobia mogą się pełnić praktycznie bez ograniczeń”²⁸⁷.

Brak transparentności był problemem w działalności X na długo przed przejściem platformy przez Elona Muska. W 2018 roku Amnesty International wielokrotnie zwracała się do firmy o opublikowanie danych dotyczących skali i charakteru nadużyć na platformie, ale prośb nie spełniono²⁸⁸. Amnesty International zaleciła następnie, aby Twitter zaczął działać w sposób transparentny w zakresie sposobu wykorzystywania uczenia maszynowego do wykrywania nadużyć oraz upublicznił dane techniczne algorytmów, na których opiera swoją działalność²⁸⁹.

W raporcie z kolejnych badań, których wyniki opublikowano w 2020 roku, Amnesty International stwierdziła, że chociaż od 2018 roku firma X poczyniła pewne postępy w rozwiązywaniu problemu przemocy i nadużyć wobec kobiet w internecie, to nadal nie wywiązuje się z obowiązków w zakresie przestrzegania praw człowieka oraz musi podjąć dalsze działania w celu ochrony praw kobiet online²⁹⁰.

5.7.1 POWSZECHNOŚĆ PRZEMOCY ZE WZGLĘDU NA PŁEĆ WSPOMAGANEJ TECHNOLOGICZNIE W SERWISIE X W 2025 ROKU

Przemoc ze względu na płeć wspomagana technologicznie na platformie X jest problemem globalnym. Analiza naukowa z 2023 roku dotycząca homofobii i transfobii w tym serwisie wykazała powszechne stosowanie języka anty-LGBTI we wszystkich siedmiu badanych językach²⁹¹. Badanie pokazało, że pomimo stosowania automatycznych systemów wykrywania mowy nienawiści na platformie X nadal pojawia się znaczna liczba treści anty-LGBTI²⁹².

Z badania ilościowego przeprowadzonego przez Amnesty International na platformie X w Polsce w 2025 roku wynika, że treści obraźliwe, w tym te o charakterze TFGBV, pozostają poważnym problemem serwisu. Analiza próby obejmującej 1387 wpisów wskazuje na to, że treści homofobiczne i transfobiczne są na X bardzo rozpowszechnione, zwłaszcza w feedach kont obserwujących polityków_polityczki, którzy nie popierają praw osób LGBTI. Przy metodologii przyjętej przez Amnesty International prawie 4% tweetów wyświetlanych w feedach takich kont miało charakter homofobiczny lub transfobiczny, a ponad 25% wszystkich tweetów związanych z LGBTI, które zobaczyły te konta, miało charakter homofobiczny lub transfobiczny (patrz: Rys. 2 poniżej). Dla porównania za homofobiczne lub transfobiczne uznano tylko 1,2% tweetów sklasyfikowanych przez konta marionetkowe z grupy „Pełne prawa” (które obserwowały polityków_polityczki popierających równouprawnienie osób LGBTI), co stanowiło 11% wszystkich treści związanych z LGBTI, które zobaczyli użytkownicy_użytkowniczki tych kont.

²⁸⁴ Tamże.

²⁸⁵ Amnesty International, *Crowdsourced Twitter study reveals shocking scale of online abuse against women*, 18 grudnia 2018, <https://www.amnesty.org/en/latest/press-release/2018/12/crowdsourced-twitter-study-reveals-shocking-scale-of-online-abuse-against-women/>.

²⁸⁶ Tamże.

²⁸⁷ Tamże.

²⁸⁸ Tamże.

²⁸⁹ Amnesty International, *Crowdsourced Twitter study reveals shocking scale of online abuse against women*.

²⁹⁰ Amnesty International, *Twitter's Scorecard: Tracking Twitter's Progress in Addressing Violence and Abuse Against Women Online*, Index: AMR 51/2993/2020, 22 września 2020, <https://www.amnesty.org/en/documents/amr51/2993/2020/en/>.

²⁹¹ Davide Locatelli i in., *A cross-lingual study of homotransphobia on Twitter*, 2023; Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), *ACL Anthology*, <https://aclanthology.org/2023.c3nlp-1.3/>.

²⁹² Davide Locatelli i in., *A cross-lingual study of homotransphobia on Twitter*.

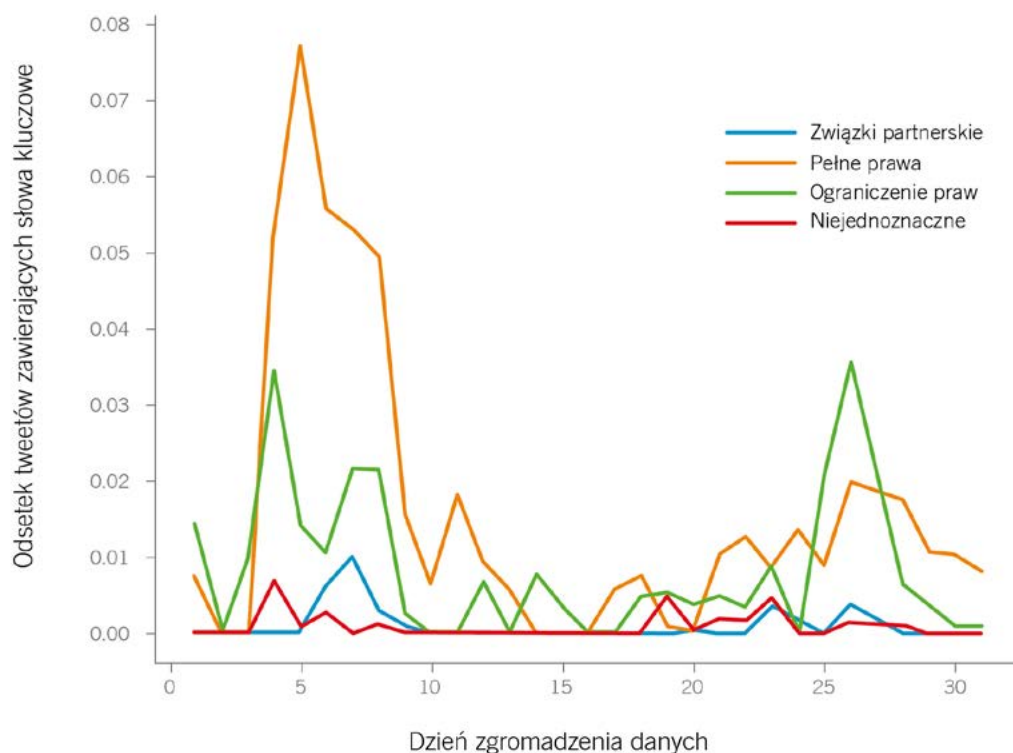
👁️ ↓ RYS. 2.

GRUPA	odsetek wszystkich wyświetlonych treści sklasyfikowanych jako homofobiczne lub transfobiczne	odsetek wyświetlonych treści związanych z osobami LGBTI, sklasyfikowanych jako homofobiczne lub transfobiczne
ZWIĄZKI PARTNERSKIE	0	0
PEŁNE PRAWA	1,2	11
OGRANICZENIE PRAW	3,7	28
NIEJEDNOZNACZNE	0,5	23

W pierwszej kolejności Amnesty International przeanalizowała rozkład procentowy wszystkich zgromadzonych tweetów pod kątem ich związku z kwestiami LGBTI. Korzystając z listy słów kluczowych²⁹³, organizacja stwierdziła, że około 0,6% ogółu tweetów wyświetlonych w feedach wszystkich 32 kont marionetkowych dotyczyło kwestii LGBTI. Chociaż przekłada się to tylko na jeden na 1700 wpisów, należy pamiętać, że jest to średnia z wszystkich treści zebranych zarówno z algorytmicznie generowanej osi czasu „Dla Ciebie”, jak i z osi czasu „Obserwowani” uporządkowanej chronologicznie. Ponieważ konta marionetkowe obserwowały wyłącznie polityków_polityczki i nie wchodziły w interakcje z żadnymi treściami na platformie, można wywnioskować, że kwestie LGBTI pozostają istotnym tematem dyskusji na X w Polsce, a przeciętny użytkownik_przeciętna użytkowniczka, którzy regularnie wchodzi w interakcje z politykami_polityczkami, przy każdym logowaniu się na X zobaczą treści związane z LGBTI.

W ramach analizy podgrup konta marionetkowe z grupy „Pełne prawa” (osoby śledzące polityków i polityczki popierających prawa osób LGBTI) odnotowały średnio dwukrotnie więcej treści związanych z LGBTI (około 1,6%) w porównaniu z trzema pozostałymi podgrupami (około 0,8%). Oznacza to, że były one najbardziej ekspozowane na treści związane z LGBTI, co ilustruje Rys. 3 umieszczony poniżej. Zjawisko to było najbardziej widoczne między piątym a dziesiątym dniem badania, co zbiegło się w czasie z przyjęciem przez polski rząd nowych przepisów dotyczących mowy nienawiści.

👁️ ↓ RYS. 3 ODSETEK TWEETÓW ZAWIERAJĄCYCH SŁOWA KLUCZOWE



²⁹³ Na liście znalazły się następujące określenia: LGBT, queer, gender, homo, gej, lesbijka, osoby tej samej płci, lesbij, parada równości, trans, lezb, lezb, płciow, płeć, płci, hetero, cispl, pedał, ciot, niebnarn, interseks, fembo.

Analiza komentarzy zebranych przez konta marionetkowe Amnesty International pod wpisami dotyczącymi kwestii LGBTI wykazała, że 26,8% z tych reakcji miało charakter homofobiczny lub transfobiczny. Co ciekawe, z analizy Amnesty International wynika, że odsetek ten był szczególnie wysoki w komentarzach i odpowiedziach na tweety wyświetlone w feedach kont z podgrupy „Pełne prawa”. Taki wynik sugeruje, że tweety popierające prawa osób LGBTI otrzymały znacznie większą liczbę homofobicznych i transfobicznych odpowiedzi niż inne tweety z badanej próby.

5.7.2 WZMOCNIENIE ALGORYTMICZNE I WYZWANIA ZWIĄZANE Z POMIAREM

Jak wyjaśniono w rozdziale „Metodologia”, badanie ilościowe Amnesty International miało ustalić, w jakim stopniu, o ile w ogóle, X zwiększa widoczność treści anty-LGBTI. Dało się to ustalić, ponieważ interfejs X umożliwia łatwe zestawienie kanałów „Dla Ciebie” i „Obserwowane”.

Jak piszemy w rozdziale „Metodologia”, badanie ilościowe opierało się na wykorzystaniu kont marionetkowych, które śledziły konta polskich polityków i polityczek na X. Konta marionetkowe nie wchodziły w interakcje z żadnymi treściami, żeby nie wzmacniać treści potencjalnie szkodliwych. W związku z tym badanie nie pozwalało na zmierzenie wpływu interakcji z takimi treściami na algorytmiczne selekcjonowanie lub wzmacnianie treści²⁹⁴. W toku niniejszego badania ilościowego, przy jego wymienionych wyżej ograniczeniach i opierając się wyłącznie na obserwacji kont, nie znaleźliśmy żadnych dowodów na algorytmiczne wzmacnianie treści związanych z osobami LGBTI lub treści anty-LGBTI.

Algorytmiczna oś czasu „Dla Ciebie” pokazuje mniej tweetów związanych z LGBTI niż oś „Obserwowani”. Można z tego wywnioskować, że politycy obserwowani przez konta marionetkowe, niezależnie od ich przynależności partyjnej czy stanowisk w kwestiach społecznych, publikowali więcej postów na temat LGBTI, niż kontom marionetkowym podsuwał algorytm, co nie świadczy jednak o istotnej amplifikacji tych treści. Ponadto w okresie objętym badaniem Sąd Najwyższy w Polsce wydał przełomowe orzeczenie znoszące wymóg udziału rodziców osób transpłciowych w postępowaniach sądowych dotyczących korekty danych w aktach stanu cywilnego: wydarzenie to mogło spowodować polityków i polityczki do publikowania wpisów na temat praw osób LGBTI.

Poniższy Rysunek 4. przedstawia wyniki eksperymentu dotyczącego osi czasu „Obserwowani” i „Dla Ciebie”. Nie zawarto w nim podziału wyników na podgrupy, ale ustalenia są spójne we wszystkich czterech grupach.

 **RYS. 4 ODSETEK TREŚCI ZWIĄZANYCH Z OSOBAMI LGBTI NA OSI CZASU „OBSERWOWANI” W PORÓWNANIU Z OSIĄ „DLA CIEBIE”**

OŚ CZASU	odsetek treści anty-LGBTI	odsetek treści sklasyfikowanych jako homofobiczne lub transfobiczne	odsetek treści pro-LGBTI
OBSERWOWANI	3,7	2,3	2,5
DLA CIEBIE	0,6	0,5	1,8

Chociaż dokładne współczynniki ważenia i wewnętrzne działanie systemu rekomendacji X pozostają nieznane, współczynniki ujawnione w publicznie dostępnym kodzie tegoż systemu (omówionym bardziej szczegółowo w rozdziale 7.) wskazują na to, że platforma nadaje wysoki priorytet treściom na osi czasu „Dla Ciebie”, które według systemu rekomendacji mogą wzbudzić zainteresowanie użytkowników i użytkowniczek, a zwłaszcza skłonić ich do napisania komentarza. Zważywszy na fakt, że konta marionetkowe nie wchodziły w interakcje z wpisami ani nie reagowały na nie, Amnesty International wysunęła roboczą hipotezę, że samo śledzenie kont o określonych poglądach na tematy polityczne lub społeczne nie wystarcza ani żeby przywabić użytkowników i użytkowniczki do treści, ani żeby zmierzyć stopień amplifikacji.

Prawdopodobnie algorytm systemu rekomendacji zacząłby w znaczącym stopniu podsuwać kontom określone treści dopiero w sytuacji, gdyby konta wchodziły w interakcje z podobnymi treściami, udostępniały

²⁹⁴ Selekcja algorytmiczna odnosi się do zautomatyzowanego procesu wybierania i organizowania treści online za pomocą algorytmów. Wzmacnianie algorytmiczne odnosi się do tendencji algorytmów do zwiększania widoczności niektórych rodzajów treści i ograniczania innych, często w oparciu o domniemane zainteresowania użytkowników platformy.

je, zostawiały polubienia lub komentowały. Biorąc pod uwagę ryzyko etyczne związane z amplifikacją szkodliwych treści, Amnesty International postanowiła nie podejmować takich działań w ramach niniejszego badania. Okoliczność ta stanowi jednak wyzwanie dla przyszłych analiz zjawiska wzmacniania algorytmicznego w serwisie X.

Nawet przy niedoskonałościach metodologii eksperymentu ilościowego badacze Amnesty International zdolali znaleźć dowody na to, że treści związane z kwestiami LGBTI są nadal powszechne na X, że znaczna część wpisów związanych z kwestiami LGBTI zawiera treści homofobiczne i transfobiczne (zarówno w postach, jak i w reakcjach na nie) oraz że użytkownicy_użytkowniczki śledzący polityków_polityczki, którzy popierają prawa osób LGBTI, są najbardziej narażeni na takie reakcje. Powszechność tych treści ma duży i negatywny wpływ na osoby LGBTI, które mieszczą się w tej kategorii. Kwestia ta zostanie omówiona bardziej szczegółowo w następnym rozdziale.

6. MIEĆ SIĘ NA BACZNOŚCI: WPŁYW PRZEMOCY ZE WZGLĘDU NA PŁEĆ WSPOMAGANEJ TECHNOLOGICZNIE NA OSOBY LGBTI

„Trudno mi o tym mówić, bo całe moje życie kręci się wokół takich sytuacji. Dlatego ciągle muszę mieć się na baczności”²⁹⁵.

Jak wykazano powyżej, X odegrał kluczową rolę w normalizacji retoryki, która dehumanizuje społeczność LGBTI w Polsce, oraz wynikającej z niej „odgórnej polaryzacji”, która trwa w polskim społeczeństwie pomimo zmiany rządu. W niniejszym rozdziale przedstawiono pięć historii osób LGBTI, które doświadczyły przemocy na tle seksualnym i przemocy w sieci – w tym gróźb przemocy, nękania online, doxingu (udostępniania w sieci w złej wierze prywatnych lub umożliwiających identyfikację informacji o konkretnej osobie) oraz ukierunkowanej nienawiści na platformie X. Osoby te opisują także, jak te doświadczenia wpłynęły na ich zdolność do swobodnego wyrażania siebie, funkcjonowanie bez dyskryminacji i poczucie bezpieczeństwa w polskim społeczeństwie. Przedstawione przypadki nie wyczerpują problemów, z jakimi borykają się osoby LGBTI w związku z TFGBV, ale dobrze ilustrują rolę platformy X w normalizacji nastrojów anti-LGBTI w Polsce oraz rozmaite skutki tego zjawiska.

6.1 HISTORIA ALEKSANDRY HERZYK

Aleksandra Herzyk jest aseksualną kobietą mieszkającą w Krakowie, drugim co do wielkości mieście w Polsce. Jest autorką komiksów i pisarką, która za pomocą mediów społecznościowych prezentuje swoje prace oraz zwiększa świadomość na temat praw człowieka i sprawiedliwości społecznej. Przestała korzystać z X po doświadczeniu TFGBV na tej platformie – na początku 2024 roku trwale usunęła swoje konto.

Aleksandra pisała na X o swojej aseksualności, w nadziei że pomoże młodszym osobom tej samej orientacji. Powiedziała Amnesty International, że reakcje na jej posty dotyczące aseksualności na X były „bardzo, ale to bardzo agresywne”²⁹⁶. Jednakże wpis, który wywołał trwającą półtora roku skierowaną wobec Aleksandry kampanię nienawiści w latach 2023–2024, dotyczył jej doświadczeń związanych z operacją zmniejszenia piersi. Aleksandra udostępniła Amnesty International przykładowe zrzuty ekranu tamtych reakcji.

Za sprawą wpisów o zabiegu niektórzy użytkownicy_użytkowniczki X zaczęli uważać Aleksandrę za transkobietę. W rezultacie Herzyk stała się celem transfobicznej nienawiści. Jak powiedziała Amnesty International, początkowo komentarze wydawały się jej zabawne, ale szybko doszło do niej, jak poważnym problemem jest mowa nienawiści:

„Z jednej strony to wszystko było śmieszne, ale wiedziałam, że skupiam na sobie hejt wymierzony w osoby transpłciowe. Czasami ludzie życzyli mi naprawdę okropnych rzeczy. Niektórzy pisali, że jeśli zobaczą mnie na siłowni lub gdziekolwiek indziej, to połamają mi kości. Albo że chcą, aby ktoś mnie zabił”²⁹⁷.

²⁹⁵ Rozmowa wideo Amnesty International z Alim (pseudonim), 26 lipca 2024.

²⁹⁶ Wywiad Amnesty International z Aleksandrą Herzyk, 27 lipca 2024.

²⁹⁷ Wywiad Amnesty International z Aleksandrą Herzyk, 27 lipca 2024.

Aleksandra zgłosiła kilka skrajnych przypadków hejtu, z jakim spotkała się na X, w tym komentarze sugerujące, że osoby transpłciowe powinny odbierać sobie życie – ale reakcje serwisu były niekonsekwentne. Niektóre posty usunięto, inne, zawierające równie groźne wezwania do przemocy, pozostały na X.

Herzyk powiedziała Amnesty International, że dopiero po opuszczeniu serwisu X zdała sobie sprawę, jaki wpływ na jej samopoczucie miało doświadczenie TFGBV:

„Byłam ciągle gotowa do dowcipnych, sarkastycznych odpowiedzi, które były bardzo nieprzyjemne. Myślałam, że dobrze się bawię, i w pewnym sensie tak było. Ale kiedy wylogowałam się z Twittera, pomyślałam: to nie była gra warta świeczki. Nie było warto poświęcać na to czasu. Wiesz co, czytasz różne wypowiedzi na swój temat, rzeczy, które nie są prawdą, ale jakoś zostają ci w głowie. To jak kara tysiąca cięć”²⁹⁸.

6.2 HISTORIA ALEGO

Ali jest 24-letnią osobą niebinarną mieszkającą w Warszawie; używa zaimków on/jego. Jest aktywistą aborcyjnym; korzysta z mediów społecznościowych, aby publikować posty na temat swojego życia jako osoby niebinarnej w Polsce. Kupuje przeznaczone dla transpłciowej młodzieży bandaż na klatkę piersiową, tak zwane bindery, oraz współpracuje z kolektywem, który udostępnia antykoncepcję awaryjną. Chociaż Ali nie korzystał z X w ramach swojej działalności aktywistycznej, większość treści spełniających znamiona TFGBV, których doświadczył, pochodziła właśnie z tej platformy.

Ali następująco opisał wpływ doświadczenia TFGBV na swoje poczucie bezpieczeństwa:

„Gdy w internecie jest więcej mowy nienawiści, ludzie stają się na nią bardziej podatni. I zazwyczaj mają więcej odwagi, aby wcielać ją w życie”²⁹⁹.

Ali zauważył, że internetowe ataki na społeczność LGBTI nasiliły się w 2019 roku:

„Myślę, że wiele aktów homofobii zaczęło się w 2019 roku z powodu kampanii prezydenckiej. Wielu polityków mówiło publicznie lub w internecie, że nie jesteśmy ludźmi, tylko jakąś ideologią. Że nie jesteśmy prawdziwymi ludźmi. Zostaliśmy zdehumanizowani. Co więcej: jeśli takie słowa padają z ust przedstawicieli władz, powstaje atmosfera przyzwolenia, ludzie myślą, że mogą tak postępować. Że to jest dozwolone. A czasami nawet nagradzane”³⁰⁰.

Ali stwierdził, że stał się celem ataków po tym, jak publicznie wypowiedział się na temat kwestii LGBTI³⁰¹. Opowiedział Amnesty International o tym, jak padł ofiarą TFGBV w 2020 roku, kiedy jako student zabrał głos podczas protestów na swojej uczelni. Protesty wybuchły po tym, jak uczelnia zaostryła zasady dotyczące używania przez studentów wybranych przez nich imion i zaimków.

„Spotkałem się z ogromną nienawiścią. Oni [osoby stosujące przemoc w internecie] znaleźli mnie również na Facebooku i Instagramie, pisali do mnie. Widziałem na Twitterze, jak zmawiali się ze sobą. Pisali: „Okej, to jest jej [sic] profil. Dobra, napiszmy coś”. A pisali, że jak pojawię się na uczelni i mnie zobaczą, to mnie zgwałcą, pobiją, zabiją. Że powinienem zrezygnować ze studiów, bo nikt nie chce tu takiego zbrodnieńca. Wszystko to znalazło się w moich prywatnych wiadomościach”³⁰².

W pewnym momencie Ali padł ofiarą doxingu, na X upubliczniono między innymi dane jego pracodawcy:

„Jacyś ludzie przyszli do mnie do pracy, wrzeszczeli na mnie, byli naprawdę agresywni”³⁰³.

Na X udostępniono również dane rodziny Alego:

„Znaleźli informacje o mojej matce i jej konto na Facebooku. Nie ujawniłem się [przed rodziną], byłem studentem, finansowo zależnym od rodziny. Ci ludzie napisali do moich rodziców: »Czy wiecie, co wasza córka [sic] robi w sieci?« . I wysłali im zrzuty ekranu z informacjami, które umieszczałem w sieci na swój temat”³⁰⁴.

²⁹⁸ Tamże.

²⁹⁹ Rozmowa wideo Amnesty International z Alim (pseudonim), 26 lipca 2024.

³⁰⁰ Tamże.

³⁰¹ Tamże.

³⁰² Tamże.

³⁰³ Tamże.

³⁰⁴ Tamże.

W czasie ataków Ali usunął konto na X, był więc zdany na informacje od przyjaciół o groźbach nadal kierowanych pod jego adresem na platformie:

„Po prostu wysyłali mi różne rzeczy i pisali: »Hej, Ali, nie wiem, czy chcesz to oglądać, ale w poście na Twitterze ktoś pisze, że chce ci zrobić krzywdę. Myślę, że powinienes o tym wiedzieć«. Miałem konto na Twitterze przez jakieś dwa miesiące, ale zalew nienawiści był po prostu przytłaczający”³⁰⁵.

Ali uważał, że musi być na bieżąco z nienawistnymi treściami:

„Czułem, że jeśli nie przeczytam tego, co piszą, nie będę gotowy na ich następny ruch. Uważałem więc, że muszę się na to [te treści] wystawić”³⁰⁶.

Ali wiedział, że jego znajomi zgłosili treści zawierające jego dane osobowe do serwisu X, ale posty nie zostały usunięte.

Opisał też uporczywe poczucie niepokoju, które towarzyszy mu od czasu tamtego doświadczenia:

„Muszę ciągle mieć się na baczności”³⁰⁷.

6.3 HISTORIA MAGDY DROPEK

Magda Dropek to 42-letnia aktywistka LGBTI, która niedawno przeprowadziła się do Warszawy z Krakowa. Od 2011 roku działa w środowiskach aktywistów LGBTI. W 2020 roku Magda zaangażowała się też w polski ruch na rzecz prawa do aborcji. Współorganizowała protesty w Krakowie i zaczęła mocniej angażować się w ruch na rzecz praw człowieka, podchodząc do nich intersekcyjnie. W lutym 2024 roku podjęła pracę w Departamencie ds. Równego Traktowania przy Kancelarii Prezesa Rady Ministrów.

Magda wyjaśniła Amnesty International, że korzysta z X przede wszystkim dlatego, że jest to ważne narzędzie w jej działalności aktywistycznej:

„Aktywizm queerowy i polityczny prowadzę głównie na Twitterze. Wiem, jak ważną funkcję komunikacyjną i aktywistyczną pełni to narzędzie... I jak ważne jest przy mówieniu o tym, co się robi i dlaczego, oraz mobilizowaniu ludzi. Bo tworzenie i organizowanie społeczności zawsze było ważnym elementem mojej działalności lokalnej”³⁰⁸.

Magda opisała X jako przestrzeń, w której społeczność LGBTI spotyka się z ogromną nienawiścią:

„Wcześniej był to raczej publiczny, czysty hejt wymierzony w osoby queer. Czytało się o sobie najgorsze rzeczy tylko dlatego, że jest się osobą queer lub aktywistką LGBTI. Jednak od dwóch lat hejt stał się bardziej osobisty. Teraz chodzi bardziej o to, jak wyglądasz, jak się prezentujesz, jak mówisz. Chodzi bardziej o to, żeby cię upokorzyć”³⁰⁹.

Według Magdy problemem na X są nie tylko treści anty-LGBTI, lecz także ataki na aktywistów tej społeczności, którzy zabierają głos w serwisie:

„Z mojego doświadczenia wynika, że jeśli piszę [lub piszemy] o tematach queer, bardzo często mój tweet ma 30, 40 polubień lub pięć komentarzy, a dzień lub kilka godzin później pojawia się około 500 nowych komentarzy [anty-LGBT]”³¹⁰.

Dropek opowiedziała Amnesty International o swoim doświadczeniu związanym z opublikowaniem tweeta wyrażającego poparcie dla ruchów Black Lives Matter i Women's Strike, który ściągnął na nią tysiące nienawistnych komentarzy, a nawet groźby telefoniczne:

„Czułam zagrożenie w realu. Zwłaszcza przez te telefony. Myślę jednak, że stałam się dzięki temu bardziej świadoma tego, co mówię, piszę oraz dlaczego to robię. Chwilami łapałam się na myśli: »a może lepiej tego nie pisać lub nie mówić«, bo moje słowa mogą zostać wykorzystane przeciwko mnie i wywołać kolejną falę nienawiści”³¹¹.

³⁰⁵ Tamże.

³⁰⁶ Tamże.

³⁰⁷ Tamże.

³⁰⁸ Wywiad Amnesty International z Magdą Dropek, 24 lipca 2024.

³⁰⁹ Tamże.

³¹⁰ Tamże.

³¹¹ Tamże.

Magda tłumaczy, że jej zdaniem uporczywe zjawisko TFGBV na platformie X nie pozwala członkom społeczności LGBTI na pełne korzystanie z prawa do wolności wypowiedzi:

„Chodzi głównie o uciszanie i pokazanie nam, że to nie jest nasze miejsce. Naszym miejscem jest milczenie, niewidoczność. A nie bycie zaangażowanym, nie podejmowanie działań, zwłaszcza aktywistycznych”³¹².

6.4 HISTORIA MAI HEBAN

Maja Heban to 34-letnia transkobieta mieszkająca w Warszawie. Od 15 lat publicznie przyznaje się do swojej transtożsamości. Aktywność w ruchu LGBTI rozpoczęła w 2019 roku, wzięła wtedy udział w proteście przeciwko homofobii w mediach społecznościowych podczas wyborów prezydenckich, w którym ludzie używali hashtagu #IMLGBT, aby zbiorowo ujawnić swoją tożsamość. Maja publikuje komentarze na temat mediów społecznościowych i praw osób transpłciowych. X to jedna z głównych platform, z których korzysta w pracy.

Maja zauważa, że poziom nienawiści na X od kilku lat jest niezwykle wysoki:

„Wydaje mi się, że nawet przed przejściem serwisu przez Elona Muska moderacja i zgłaszanie prawie w ogóle nie działały. Pamiętam, że wiele razy robiłam raban, zgłaszałam wyjątkowo negatywne, pełne nienawiści treści. Nie podejmowano żadnych działań, nie dało się usunąć tych komentarzy. Nie chodziło nawet o blokowanie użytkowników, ale konkretnych tweetów o tym, że powinnam się zabić”³¹³.

Maja wyjaśnia Amnesty International, że X odgrywa w Polsce rolę ważnej platformy politycznej, w rezultacie czego nadal korzysta z niej wielu aktywistów_aktywistek, w tym ona sama, pomimo że doznaje tam przemocy ze względu na płeć (TFGBV).

Maja uważa, że X tworzy środowisko, w którym coraz bardziej dopuszczalne są nastroje anty-LGBTI. Opisała niektóre treści skierowane przeciwko społeczności LGBTI:

„Można powiedzieć wszystko – porównywać osoby LGBT do zwierząt, gwałcicieli, pedofilów. Wszystko uchodzi”³¹⁴.

Maja stwierdziła, że ogrom nienawiści, z jaką się spotkała, w pewnym stopniu ją znieczulił:

„Ponieważ tak otwarcie mówię o sobie, wszelką nienawiść, obraźliwe wypowiedzi, misgenderowanie czy deadnaming traktuję po prostu jako część życia. Od czasu do czasu zastanawiam się: przecież niektórzy ludzie nie żyją w ten sposób, prawda? Być może otwarcie mówią o swojej transtożsamości, ale zarazem nie spodziewają się, że codziennie będą słyszeć o sobie taki hejt. Nie spodziewają się, że obcy ludzie będą im codziennie mówić, żeby się zabili. Może nie powinnam się do tego przyzwyczajać”³¹⁵.

Maja opisała Amnesty International szczególnie rażący przykład transfobii na platformie X: w 2022 roku zdjęcia z grupy wsparcia osób trans na Facebooku zostały opublikowane na platformie X przez aktywistkę antytrans:

„Wkradł się [do grupy] i pobrał zdjęcia wyników mastektomii młodego transmężczyzny, który pozował z bliznami po operacji. Aktywista zanonimizował zdjęcia, ale w zasadzie opublikował je na Twitterze. Napisał, że »to 14-letnia dziewczyna, której odcięto piersi i okaleczono«. Fotografie obejrzano setki tysięcy razy”³¹⁶.

6.5 HISTORIA NATHANA BRYZY

Nathan Bryza to 21-letnia niebinarna osoba transpłciowa, używająca męskich zaimków, mieszkająca we Wrocławiu. Pracuje w firmie transportowej.

Nathan opisał, jak TFGBV na X wpłynęło na jego poczucie bezpieczeństwa:

„Twitter to specyficzne miejsce, takie piekielko. Ostatnio, przy okazji rozpoczęcia igrzysk olimpijskich, sytuacja jeszcze się pogorszyła: pojawiło się dużo nienawiści wobec osób transpłciowych, mimo że

³¹² Tamże.

³¹³ Wywiad Amnesty International z Mają Heban, 30 lipca 2024.

³¹⁴ Tamże.

³¹⁵ Wywiad Amnesty International z Mają Heban, 30 lipca 2024.

³¹⁶ Tamże.

w zawodach bierze udział kobieta, która nie jest trans. Narasta panika, przez którą boję się, że ludzie zobaczą, że jestem trans i zaatakują mnie na ulicy”³¹⁷.

Nathan opisał typ nienawiści wobec osób LGBTI, z którą regularnie spotyka się na X:

„[Piszą o nas:] nienormalni ludzie, zepsuci, chorzy. Myślę, że »chore« to najczęściej używane słowo. I jeszcze że jesteśmy grzesznikami. Że jesteśmy lubieżni. Że próbujemy zrobić z siebie kogoś wyjątkowego”³¹⁸.

TFGBV sprawia, że Nathan nie czuje się bezpiecznie w świecie offline:

„Niedawno zmieniłem pracodawcę i codziennie myślę sobie w pracy: »Kiedy w końcu ci ludzie dowiedzą się, że jestem queer?«. Kiedy patrzą na mnie, wydaje mi się, że coś podejrzewają, ale nie mówię o tym otwarcie, i boję się, że pewnego dnia przyjdę do pracy i usłyszę od nich to, co czytam w sieci na temat siebie i znajomych. To przerażające, bo czuję się naprawdę zagrożony”³¹⁹.

Doświadczenie Nathana, który nie może w pełni wyrażać swojej tożsamości w pracy, jest wśród polskiej społeczności LGBTI powszechne. Jak wynika z badań naukowych, około 35% pracujących osób LGBTI uważa, że musi ukrywać swoją tożsamość w miejscu pracy z obawy przed dyskryminacją³²⁰.

Nathan wyjaśnił, jak bardzo wpłynęło to na jego emocje:

„Jest mi z tego powodu naprawdę smutno. Czuję złość, bo bez względu na to, jak bardzo się bronimy, nikt nam nie wierzy. To mnie głęboko przygnębia”³²¹.

Nathan mówił również o braku skutecznej moderacji treści w języku polskim i stwierdził, że pogłębia to problem hejtu w internecie:

„O ile mi wiadomo, na polskim Twitterze jest tylko jeden moderator. Nie widzę, żeby jakiegokolwiek treści były usuwane. Nie wiem, ile zgłoszeń przekazuje się wyżej. Wszystko nadal wisi, bez względu na to, jak bardzo jest nienawistne”³²².

Opisane w niniejszym rozdziale doświadczenia społeczności LGBTI w Polsce pokazują szkody, do których przyczyniła się firma X poprzez brak odpowiedniej reakcji na TFGBV na swojej platformie. Powszechność aktów TFGBV na platformie X sprawiła, że członkowie społeczności LGBTI w Polsce czują się zagrożeni zarówno w sieci, jak i poza nią, co negatywnie wpływa na ich zdrowie psychiczne i możliwość swobodnego wyrażania siebie. Narażenie na niemal ciągły kontakt ze szkodliwymi treściami na X poważnie narusza ich prawo do wolności wypowiedzi, niedyskryminacji i życia wolnego od przemocy na tle płciowym. Przedstawione powyżej studia przypadków jednoznacznie dowodzą, że X nie wywiązuje się ze swoich obowiązków w zakresie poszanowania praw człowieka osób LGBTI korzystających z platformy w Polsce, a także nie podjął odpowiednich działań w celu ograniczenia ryzyka wystąpienia przemocy ze względu płeć, co jest wymogiem DSA.

W następnym rozdziale skupimy się na tym, w jaki sposób model biznesowy X może przyczyniać się do dalszego rozprzestrzeniania treści anty-LGBTI na tej platformie.

³¹⁷ Rozmowa wideo Amnesty International z Nathanem Bryzą, 9 sierpnia 2024.

³¹⁸ Tamże.

³¹⁹ Tamże.

³²⁰ United Nations Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity, Country visit to Poland, 18–29 listopada 2024, *End of mission statement*, akapit 38.

³²¹ Rozmowa wideo Amnesty International z Nathanem Bryzą, 9 sierpnia 2024.

³²² Tamże.

7. BIZNES NIENAWIŚCI: JAK MODEL BIZNESOWY X ZWIĘKSZA RYZYKO NARUSZANIA PRAW CZŁOWIEKA I KRZYWDZI LUDZI

„Nie jest zbyt przyjaznym miejscem, czytanie tam różnych rzeczy powoduje wielką frustrację. Bardzo trudno jest zachować zdrowie psychiczne, gdy korzysta się z Twittera”³²³.

7.1 MODEL BIZNESOWY OPARTY NA INWIGILACJI

Amnesty International już wcześniej ustaliła, że firmy technologiczne Meta (spółka macierzysta Facebooka) i Google stosują model biznesowy oparty na inwigilacji, który polega na ciągłym gromadzeniu danych od użytkowników_użytkowniczek w celu lepszego dopasowania reklam wyświetlanych na platformie. Jest to podejście z gruntu sprzeczne z prawem do prywatności i stanowi zagrożenie dla szeregu praw człowieka, w tym wolności opinii i wypowiedzi, wolności myśli oraz prawa do równości i niedyskryminacji³²⁴.

Również Federalna Komisja Handlu (FTC) stwierdziła, że główne serwisy społecznościowe i streamingowe, w tym X, prowadzą szeroko zakrojony nadzór nad konsumentami w celu monetyzacji ich danych osobowych, nie zapewniając im jednocześnie właściwej ochrony online³²⁵.

Niniejsza część raportu zawiera opis cech modelu biznesowego X opartego na inwigilacji oraz wyjaśnienie, dlaczego model ten w sposób systemowy zagraża prawom człowieka.

7.1.1 ZALEŻNOŚĆ OD DANYCH UŻYTKOWNIKÓW_UŻYTKOWNICZEK

Model biznesowy X bazuje na gromadzeniu danych użytkowników_użytkowniczek na każdym etapie korzystania przez nich z platformy oraz w sposób, którego nie da się uznać za zgodny z obowiązkiem firmy

³²³ Wywiad Amnesty International z Mateuszem Kaczmarkiem, 28 lipca 2024.

³²⁴ Amnesty International, *Surveillance Giants: How the Business Model of Google and Facebook Threatens Human Rights*, Index: POL 30/1404/2019, 21 listopada 2019, <https://www.amnesty.org/en/documents/pol30/1404/2019/en/>.

³²⁵ FTC, *FTC staff report finds large social media and video streaming companies have engaged in vast surveillance of users with lax privacy controls and inadequate safeguards for kids and teens*, 19 września 2024, <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-staff-report-finds-large-social-media-video-streaming-companies-have-engaged-vast-surveillance>.

do poszanowania prawa do prywatności³²⁶. Dane tego typu mają kluczowe znaczenie dla X, ponieważ pomagają platformie przewidywać, które treści będą interesować użytkowników_użytkowniczkę, co w dużej mierze decyduje o wartości X dla reklamodawców³²⁷.

Polityka prywatności X stanowi, że w celu otwarcia konta należy przekazać firmie pewne informacje, co oznacza, że gromadzenie danych jest kluczowym wymogiem dostępu do produktów i usług platformy³²⁸. Przy rejestracji wymagane jest podanie imienia i nazwiska, nazwy użytkownika, hasła, adresu e-mail i numeru telefonu, daty urodzenia, języka wyświetlania oraz informacji potrzebnych do jednoetapowego logowania za pośrednictwem serwisu zewnętrznego³²⁹. Użytkownicy_użytkowniczki platformy mogą również zdecydować się na udostępnienie swojej lokalizacji w profilu i postach oraz na przesłanie swojej książki adresowej w celu identyfikacji znajomych na X³³⁰.

W polityce prywatności X zaznaczono, że zbierane są też dane o ustawieniach preferencji, a także inne informacje na temat tego, jak użytkownicy_użytkowniczki korzystają z platformy: „Gdy korzystasz z naszych usług, zbieramy informacje o tym, jak korzystasz z naszych produktów i usług. Używamy tych informacji, żeby dostarczać Ci produkty i usługi, pomagać w zapewnieniu bezpieczeństwa na platformie i szacunku dla wszystkich użytkowników X oraz dostosowywać treści **do Twoich potrzeb**”³³¹. Nacisk na trafność proponowanych treści świadczy o kluczowym znaczeniu roli przyciągania uwagi użytkowników_użytkowniczek dla modelu biznesowego X; firma gromadzi dane o użytkownikach_użytkowniczkach, aby rekomendować im treści, które sprawią, że pozostaną na platformie dłużej, co z kolei pozwala X gromadzić więcej danych na ich temat. Ta szeroko zakrojona inwigilacja ze strony korporacji kłóci się z prawem do prywatności i może ograniczać prawo do wolności opinii i wypowiedzi oraz prawo do niedyskryminacji.

X wyraźnie zastrzega również, że dane będą gromadzone wyłącznie na potrzeby rekomendacji dotyczących ofert pracy i reklam, deklarując, że będzie gromadzić i wykorzystywać dane osobowe (takie jak historia zatrudnienia, wykształcenie, preferencje dotyczące zatrudnienia, umiejętności i kompetencje, aktywność w poszukiwaniu pracy i zaangażowanie „itp.”) w celu rekomendowania stanowisk, umożliwienia pracodawcom identyfikacji potencjalnych kandydatów oraz wyświetlania bardziej trafnych reklam ukierunkowanych³³².

W 2022 roku FTC podjęła działania przeciwko X za nieuczciwe wykorzystanie danych dotyczących bezpieczeństwa kont do celów reklamy ukierunkowanej, co poskutkowało nałożeniem na firmę kary w wysokości 150 milionów dolarów oraz objęciem jej trwałym zakazem czerpania korzyści z danych zebranych w nieuczciwy sposób³³³.

7.1.2 REKLAMY UKIERUNKOWANE

Od 2013 roku prawie wszystkie przychody X pochodziły z reklam ukierunkowanych, umieszczanych w serwisie³³⁴. W 2021 roku reklamy stanowiły ponad 90% przychodów firmy, wynoszących łącznie 5,1 miliardów dolarów³³⁵.

Już w 2023 roku było oczywiste, że reklamy pozostają głównym źródłem dochodów X. Serwis odnotował 40-procentowy spadek przychodów po tym, jak ponad 500 reklamodawców wstrzymało nakłady z powodu

³²⁶ Cornelius Puschmann i Jean Burgess, *The politics of Twitter data*, 23 January 2013, HIIG Discussion Paper Series, styczeń 2013, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2206225.

³²⁷ FTC, *FTC staff report finds large social media and video streaming companies have engaged in vast surveillance of users with lax privacy controls and inadequate safeguards for kids and teens*.

³²⁸ X, *Privacy policy*, <https://x.com/en/privacy#update>, dostęp: 5 lipca 2025.

³²⁹ Tamże.

³³⁰ Tamże.

³³¹ Tamże, podkreślenie Amnesty International.

³³² Tamże.

³³³ FTC, *A look behind the screens: examining the data practices of social media and video streaming services*, wrzesień 2024, https://www.ftc.gov/system/files/ftc_gov/pdf/Social-Media-6b-Report-9-11-2024.pdf.

³³⁴ BBC News, *How does Twitter make money?*, 7 listopada 2013, <https://www.bbc.co.uk/news/business-24397472>.

³³⁵ Bloomberg UK, *Documents show how Musk's X plans to become the next Venmo*, 18 czerwca 2024, <https://www.bloomberg.com/news/articles/2024-06-18/documents-show-how-musk-s-x-plans-to-become-the-next-venmo>; The Guardian, *Twitter hit by 40% revenue drop amid ad squeeze*, 18 stycznia 2023, <https://www.theguardian.com/technology/2023/jan/18/twitter-revenue-drop-advertising-squeeze-elon-musk>; Mashable, *Elon Musk's X revenue has officially plummeted*, 18 czerwca 2024, <https://mashable.com/article/twitter-x-revenue-falls-x-payments-plans>.

obaw związanych ze zmianami wprowadzonymi w polityce X³³⁶. Warunki korzystania z serwisu, ostatni raz zaktualizowane 15 listopada 2024 roku, jasno określają kluczową rolę reklam wyświetlanych na platformie: „**Na platformie zobaczysz reklamy:** w zamian za dostęp do Usług X oraz nasi zewnętrzni dostawcy i reklamodawcy mogą wyświetlać użytkownikom_użytkowniczkom treści reklamowe”³³⁷.

Istnieją trzy główne sposoby reklamowania się na X: promowanie tweeta, który pojawi się na osi czasu innych użytkowników_użytkowniczek, promowanie całego konta lub promowanie popularnego tematu³³⁸. Podobnie jak wiele innych mediów społecznościowych X zazwyczaj pobiera opłaty od reklamodawców w zależności od liczby interakcji wygenerowanych przez ich treści – reklamodawcy płacą za kliknięcie lub retweet³³⁹, a platforma gromadzi jak największą ilość danych użytkowników_użytkowniczek w celu trafnego targetowania reklam, co z kolei przekłada się na liczbę kliknięć lub retweetów. X posiada również „system licytacji”: reklamodawcy konkurują o to, aby ich treści pojawiały się w określonej przestrzeni na platformie³⁴⁰.

W momencie tworzenia niniejszego raportu X nie jest już notowany na giełdzie, co utrudnia uzyskanie aktualnych informacji na temat źródeł przychodów firmy³⁴¹. Większość informacji dotyczących jej przychodów, w tym problemów z nimi związanych, pochodzi z wewnętrznych przecieków, a nie z oficjalnych źródeł³⁴². Według nieoficjalnych doniesień w pierwszych sześciu miesiącach 2023 roku przychody X spadły o prawie 40% w porównaniu z tym samym okresem roku poprzedniego, a w pierwszym kwartale 2023 roku firma straciła 456 milionów dolarów³⁴³.

7.1.3 ALTERNATYWNE ŹRÓDŁA PRZYCHODÓW

Od chwili przejścia firmy w 2022 roku Elon Musk starał się zmodyfikować model biznesowy X, tak aby stworzyć źródła przychodów niezależne od reklam. Obejmowały one plan subskrypcji X Premium oraz usługę subskrypcji dla twórców³⁴⁴. Jednak żadne z nowych rozwiązań nie zniwelowało spadku przychodów spowodowanego gwałtownym odpływem reklamodawców³⁴⁵.

W ramach ambitnych planów Muska zakładających rozbudowę platformy do „aplikacji do wszystkiego”³⁴⁶ X starał się też uzyskać licencję na świadczenie usług płatniczych o nazwie X Payments. Jednak według wewnętrznych dokumentów firma planuje wykorzystywać tę usługę głównie w celu zwiększenia zaangażowania i aktywności na platformie. Serwis zakłada, że X Payments mają nie pobierać opłat za większość usług³⁴⁷, co wskazuje na to, że pomimo pozornie znaczących zmian w modelu biznesowym X nadal będzie się koncentrować na generowaniu zaangażowania użytkowników_użytkowniczek.

X PREMIUM

X Premium to dobrowolna, płatna subskrypcja, która według X oferuje użytkownikom_użytkowniczkom dodatkowe funkcje poprawiające komfort korzystania z platformy poprzez podniesienie jakości rozmów³⁴⁸. W ramach X Premium dostępne są trzy poziomy: Basic, Premium i Premium+³⁴⁹. Każdy zapewnia użytkownikom_użytkowniczkom korzystanie z bardziej zaawansowanego algorytmu (priorytetyzacja

³³⁶ The Guardian, *Twitter hit by 40% revenue drop amid ad squeeze*; Mashable, *Elon Musk's X revenue has officially plummeted*.

³³⁷ X, *Terms of Service*, 15 listopada 2024, <https://x.com/en/tos>.

³³⁸ BBC News, *How does Twitter make money?*.

³³⁹ Tamże.

³⁴⁰ Tamże.

³⁴¹ Mashable, *Elon Musk's X revenue has officially plummeted*.

³⁴² Tamże.

³⁴³ Tamże.

³⁴⁴ Mashable, *Elon Musk's X revenue has officially plummeted*.

³⁴⁵ Tamże.

³⁴⁶ Bloomberg UK, *Documents show how Musk's X plans to become the next Venmo*; Mashable, *Elon Musk's X revenue has officially plummeted*.

³⁴⁷ Mashable, *Elon Musk's X revenue has officially plummeted*.

³⁴⁸ X, *About X Premium*, <https://help.x.com/en/using-x/x-premium>, dostęp: 2 lipca 2025.

³⁴⁹ Tamże.

odpowiedzi – komentarze użytkownika_użytkowniczki z danego poziomu są bardziej widoczne na platformie), a także dodatkowe narzędzia do tworzenia treści³⁵⁰.

Poziom Basic zapewnia dodatkowe funkcje, takie jak edycja postów, większa liczba znaków w poście i dłuższe filmy, priorytetyzacja odpowiedzi, formatowanie tekstu, foldery zakładki i niestandardowe ikony aplikacji³⁵¹.

Poziomy Premium zapewniają wszystkie powyższe funkcje, a także „niebieski znaczek” (wcześniej będący symbolem weryfikacji), mniej reklam, dostęp do podziału zysków z reklam, subskrypcję twórców, wyższy priorytet odpowiedzi, weryfikację tożsamości, dostęp do studia multimedialnego oraz do Groka, czyli chatbota stworzonego przez xAI dzięki generatywnej AI³⁵².

Premium+ obejmuje wszystkie funkcje Premium, a także dodatkowe korzyści, takie jak brak reklam w całym serwisie i najwyższy priorytet odpowiedzi³⁵³.

Maja Heban podzieliła się z Amnesty International swoją opinią, że za sprawą funkcji Premium X stał się mniej bezpieczny:

„Obecny model monetyzacji – w którym można zapłacić za weryfikację konta, a następnie otrzymywać wynagrodzenie za angażowanie użytkowników – oznacza, że... ludzie są zachęceni do zwiększania zaangażowania, nawet jeśli oznacza to fabrykowanie informacji, szerzenie strachu, rozpowszechnianie fałszywych wiadomości, nękanie... Dopóki ludzie odpowiadają i zarzucają ci kłamstwo, dopóty coś zyskujesz. A więc w pewnym sensie model zachęca do szerzenia dezinformacji i mowy nienawiści”³⁵⁴.

7.2 ALGORYTMY OPARTE NA ZAANGAŻOWANIU I ARCHITEKTURA SYSTEMU REKOMENDACJI X

W niniejszym podrozdziale analizujemy działanie systemu rekomendacji X, przedstawiając wagi przypisane różnym interakcjom, które użytkownicy_użytkowniczki mogą podejmować na platformie. Celem jest ustalenie, w jaki sposób platforma zwiększa zaangażowanie i personalizację. Analiza systemu rekomendacji wykazała, że starannie zaprojektowane zabezpieczenia, wzmocnione autentycznym zaangażowaniem społeczności, mogłyby znacznie ograniczyć potencjalne szkody wyrządzane przez system. Wydaje się jednak, że dla X priorytetem są wskaźniki zaangażowania, zaś wspomniane zabezpieczenia są albo wdrażane w niewystarczającym stopniu, albo całkowicie pomijane.

Modele biznesowe oparte na inwigilacji zazwyczaj przedkładają maksymalizację zaangażowania użytkowników_użytkowniczek nad wszystkie inne czynniki; im dłużej ktoś pozostaje na platformie, tym więcej danych można o nim zebrać i tym precyzyjniej można kierować do niego reklamy³⁵⁵. Amnesty International stwierdziła we wcześniejszym raporcie, że taki model biznesowy może prowadzić do tego, iż algorytmy rekomendujące zaczynają promować treści, które podlegają do nienawiści, dyskryminują i dzielą, jako że właśnie takie treści często najbardziej przyciągają użytkowników_użytkowniczki platformy³⁵⁶.

Algorytmiczne promowanie treści częściowo wynika ze spersonalizowanych rekomendacji. W serwisie X spersonalizowane rekomendacje dotyczą tweetów, wydarzeń, tematów, hashtagów i użytkowników_użytkowniczek³⁵⁷.

Platforma X posiada dwie osie czasu: „Obserwowane” i „Dla Ciebie”. Algorytm rekomendacji koncentruje

³⁵⁰ House of Commons Science Innovation and Technology Committee, *Oral evidence: Social media, misinformation and harmful algorithms*, HC 441.

³⁵¹ X, *About X Premium*.

³⁵² Tamże.

³⁵³ Tamże.

³⁵⁴ Wywiad Amnesty International z Mają Heban, 30 lipca 2024.

³⁵⁵ Amnesty International, *Surveillance Giants*.

³⁵⁶ Amnesty International: *The Social Atrocity: Meta and the Right to Remedy for the Rohingya*, Index: ASA 16/5933/2022, 28 września 2022, <https://www.amnesty.org/en/documents/asa16/5933/2022/en/>; Amnesty International, „A Death Sentence for My Father”: *Meta’s Contribution to Human Rights Abuses in Northern Ethiopia*, Index: AFR 25/7292/2023, 31 października 2023, <https://www.amnesty.org/en/documents/afr25/7292/2023/en/>.

³⁵⁷ Kayla Duskin i in., „Echo chambers in the age of algorithms: an audit of Twitter’s friend recommender system”, may 2024, *WEBSCI24: Proceedings of the 16th ACM Web Science Conference*, <https://dl.acm.org/doi/abs/10.1145/3614419.3643996>.

się głównie na tej drugiej – ma ona pokazywać użytkownikom_ użytkowniczkom nowe treści z kont, których jeszcze nie obserwują, a także treści z kont już przez nich obserwowanych. Uchodzi za główny kanał platformy³⁵⁸. Oś czasu „Dla Ciebie” wprowadzono w styczniu 2023 roku po przebudowie serwisu³⁵⁹. Pierwotnie kanały X wyświetlały tweety z kont obserwowanych przez użytkownika_ użytkowniczkę w porządku chronologicznym, a następnie posty polubione lub komentowane przez obserwowane konta³⁶⁰. Przed 2022 rokiem X zaczął wyświetlać rekomendacje postów pod hasłem „Może Ci się spodobać”. Oś „Dla Ciebie” bazuje na tym samym modelu zaangażowania, lecz stanowi odejście od układu chronologicznego³⁶¹. X domyślnie wyświetla teraz oś czasu „Dla Ciebie”³⁶².

Podstawą algorytmicznego systemu rekomendacji X jest zbiór modeli i funkcji, które wyodrębniają dane ukryte w tweetach, danych użytkowników_ użytkowniczek i danych dotyczących zaangażowania³⁶³.

W publicznie dostępnym poście na blogu X z 2023 roku serwis opisuje ten model jako próbę udzielenia odpowiedzi na pytania: „Jakie jest prawdopodobieństwo, że w przyszłości nawiążesz interakcję z innym użytkownikiem_inną użytkowniczką?” lub „Jakie społeczności istnieją na Twitterze i jakie tweety są w nich popularne?”³⁶⁴. Szczegółowe informacje i analizy dotyczące architektury systemu rekomendacji X pochodzą ze wspomnianego wpisu na blogu oraz z przeprowadzonej przez Amnesty International analizy elementów kodu źródłowego platformy, które zostały upublicznione przez X w 2023 roku³⁶⁵.

Proces rekomendacji ma trzy główne cechy³⁶⁶:

- ❑ wyszukiwanie kandydatów (pobieranie najbardziej angażujących tweetów z różnych źródeł rekomendacji);
- ❑ ocena każdego wytypowanego tweeta w celu przypisania mu wskaźnika prawdopodobieństwa, że jego treść zaangażuje użytkownika_ użytkowniczkę. Model przewiduje prawdopodobieństwo różnych interakcji, także to, czy użytkownik_ użytkowniczka polubi tweet, udostępni go, odpowie na niego, kliknie go, a nawet oznaczy jako nieodpowiedni³⁶⁷;
- ❑ zastosowanie metod heurystycznych i filtrów, na przykład odsiewanie tweetów od zablokowanych użytkowników_ użytkowniczek, treści nienadających się do wyświetlania w pracy oraz tweetów, które zostały już wyświetlone.

Oś czasu „Dla Ciebie” jest kształtowana poprzez sprzężenie tych trzech funkcji, a następnie zastosowanie logiki wzmacniającej (zwiększającej widoczność określonych tweetów). Całość usługi funkcjonuje pod nazwą Home Mixer³⁶⁸. Proces jest przeprowadzany około 5 miliardów razy dziennie i trwa średnio poniżej 1,5 sekundy, dzięki czemu każdego dnia na urządzenia użytkowników_ użytkowniczek serwisu trafia 150 miliardów tweetów³⁶⁹.

Cały proces wraz z wszystkimi szczegółami technicznymi i zastosowanymi metodami przedstawiono graficznie na Rysunku 5.

³⁵⁸ X, *Twitter's recommendation algorithm*, 31 marca 2023, https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm.

³⁵⁹ Washington Post, *Elon Musk's Twitter pushes hate speech, extremist content into „For You” pages*, 30 marca 2023, <https://www.washingtonpost.com/technology/2023/03/30/elon-musk-twitter-hate-speech/>.

³⁶⁰ Tamże.

³⁶¹ Tamże.

³⁶² Tamże.

³⁶³ X, *Twitter's recommendation algorithm*.

³⁶⁴ Tamże.

³⁶⁵ Zobacz: <https://github.com/twitter/the-algorithm>.

³⁶⁶ X, *Twitter's recommendation algorithm*.

³⁶⁷ Odbywa się to przy użyciu głębokiej sieci neuronowej, a modelu tego nie udostępniono na licencji *open source*. Zob. Kevin Feng i in., *Probing the ethical boundaries of personalization: a case study of Twitter's recommendation algorithm*, 2024, CSE 581 – Computing Ethics, https://homes.cs.washington.edu/~micibr/assets/pdf/ethical_personalization_paper.pdf.

³⁶⁸ X, *Twitter's recommendation algorithm*.

³⁶⁹ Tamże.

👁️ ↓ RYS. 5: GRAFICZNA PREZENTACJA PRZEPŁYWU DANYCH W USŁUDZE HOME MIXER, KTÓRA GENERUJE KANAŁ „DLA CIEBIE”



Aby wygenerować spersonalizowany kanał, system rekomendacji musi najpierw pobrać pulę wytypowanych tweetów, które mogłyby być interesujące dla użytkownika_użytkowniczki. X selekcjonuje tweety z dwóch głównych obszarów: treści w sieci użytkownika_użytkowniczki (tweety z kont obserwowanych przez użytkownika_użytkowniczkę) oraz treści spoza sieci użytkownika_użytkowniczki (tweety z innych kont)³⁷⁰. Średnio system typuje około 1500 tweetów dla danego użytkownika_danej użytkowniczki³⁷¹, po około połowie z każdej kategorii³⁷². Dzięki temu na osi czasu „Dla Ciebie” pojawiają się treści zarówno znane, jak i zupełnie nowe.

Po wytypowaniu puli tweetów system rekomendacji X wykorzystuje zestaw algorytmów rankingowych bazujących na uczeniu maszynowym³⁷³, aby ocenić użyteczność wpisów pod kątem danego użytkownika_danej użytkowniczki. Ten etap stanowi istotę personalizacji, w przypadku której wielkoskalowy model sieci neuronowej przewiduje reakcję użytkownika_użytkowniczki na każdy tweet i przypisuje tweetom odpowiednią ocenę trafności.

W 2023 do tworzenia rankingów służyła głęboka sieć neuronowa obejmująca około 48 milionów parametrów. Model jest stale szkolony na podstawie ogromnych dzienników interakcji platformy: uczy się na podstawie zbiorczych zachowań użytkowników_użytkowniczek X w czasie zbliżonym do rzeczywistego. Ilekroć użytkownicy wchodzi w interakcje z tweetami (lub je ignorują), dostarczają modelowi dane szkoleniowe na temat tego, jakie treści są popularne wśród poszczególnych odbiorców_odbiorczyń.

Model korzysta z tysięcy parametrów wejściowych, obejmujących wszystkie cechy użytkownika_użytkowniczki, tweeta i ich zależność³⁷⁴. Cechy użytkownika_użytkowniczki to dane demograficzne, domniemane zainteresowania i dotychczasowa aktywność. Cechy tweeta to osadzenie tekstu, autorstwo,

³⁷⁰ Aneesh Sharma i in., *GraphJet: real-time content recommendations at Twitter*, 2016, *Proceedings of the VLDB Endowment*, Volume 9, Issue 13, <https://www.vldb.org/pvldb/vol9/p1281-sharma.pdf>.

³⁷¹ Dotyczy to każdego żądanego posta – każdej treści, która pojawia się na osi „Dla Ciebie” użytkownika.

³⁷² X, *Twitter's recommendation algorithm*.

³⁷³ Model niedostępny w kodzie *open source*.

³⁷⁴ Anthony Alford, *Twitter open-sources recommendation algorithm*, 11 kwietnia 2023, <https://www.infoq.com/news/2023/04/twitter-algorithm/>.

statystyki zaangażowania i wskaźniki społeczności. Cechy zależności uwzględniają to, czy użytkownik_użytkowniczka obserwuje autora_autorkę, jak często wchodził_wchodziła w interakcje z podobnymi tweetami oraz czy tweet został polecany przez znajomego_znajomą. Na podstawie opisanych powyżej danych wejściowych sieć neuronowa generuje zestaw przewidywanych wyników zaangażowania, na przykład prawdopodobieństwo, że użytkownik_użytkowniczka polubi tweet, udostępni go, odpowie na niego, kliknie go czy nawet oznaczy jako nieodpowiedni³⁷⁵.

Według serwisu X model jest wielozadaniowym systemem uczącym się, który generuje około 10 prognoz dla każdego tweeta (każda odpowiada konkretnemu działaniu użytkownika_użytkowniczki)³⁷⁶. Aby przekształcić te prognozy w konkretny wynik, system X stosuje zaprogramowaną formułę ważoną, która nadaje priorytet niektórym działaniom.

Chociaż wersja modelu rankingowego jest dostępna na licencji *open source* (dotyczy to także jego architektury i hiperparametrów oraz przykładowego procesu szkolenia), wagi stosowane w modelu nie zostały udostępnione. X uzasadnił to względami prywatności – upubliczniony model mógłby zostać ponownie przeszkolony na danych publicznych lub częściowo zrandomizowany³⁷⁷. Mimo to możemy wyciągać wnioski z publicznie dostępnych wag, które szczegółowo przedstawiono poniżej (Rys. 6).

 **RYS. 6 WAGI DLA KAŻDEJ PRZEWIDYWANEJ MOŻLIWOŚCI**

DZIAŁANIE	WAGA	OPIS
POLUBIENIE	0,5	Przewidywane prawdopodobieństwo, że użytkownik_użytkowniczka polubi (doda do ulubionych) tweet: bardzo mały wpływ na ocenę końcową w ranking.
RETWEET	1,0	Przewidywane prawdopodobieństwo, że użytkownik_użytkowniczka udostępni tweet: słaby wpływ, nieznacznie silniejszy niż polubienie.
ODPOWIEDŹ	13,5	Przewidywane prawdopodobieństwo, że użytkownik_użytkowniczka odpowie: silnie zwiększa popularność tweetów, które prowadzą do bezpośredniej rozmowy użytkowników_użytkowniczek.
DOBRE KLIKNIĘCIE W PROFILL	12,0	Prawdopodobieństwo, że użytkownik_użytkowniczka kliknie profil autora_autorki, a następnie polubi/odpowie: ma wartość prawie tak samo wysoką jak odpowiedź.
ODTWARZANIE WIDEO >50%	0,005	Prawdopodobieństwo, że użytkownik_użytkowniczka obejrzy ponad 50% materiału wideo: praktycznie zerowy wpływ na ranking.
ODPOWIEDŹ I INTERAKCJA Z AUTOREM_AUTORKĄ	75,0	Prawdopodobieństwo, że użytkownik_użytkowniczka odpowie, a autor_autorka wejdzie w interakcję: najwyższa nagroda za podtrzymanie rozmowy.
DOBRE KLIKNIĘCIE (OTWARCIE ROZMOWY)	11,0	Prawdopodobieństwo, że użytkownik_użytkowniczka otworzy widok rozmowy, a następnie polubi/odpowie: świadczy o głębokim zainteresowaniu rozmową.
DOBRE KLIKNIĘCIE V2 (2 MIN W TRYBIE ROZMOWY)	10,0	Prawdopodobieństwo, że użytkownik_użytkowniczka pozostanie w widoku rozmowy dłużej niż dwie minuty: mocny wskaźnik siły zaangażowania.
NEGATYWNY FEEDBACK	-74,0	Prawdopodobieństwo wystawienia negatywnej opinii (np. „Pokaż mniej”, blokada lub wyciszenie): silnie ogranicza niepopularne lub niechciane treści.

³⁷⁵ Kevin Feng i in., *Probing the ethical boundaries of personalization: a case study of Twitter's recommendation algorithm*.

³⁷⁶ X, *Twitter's recommendation algorithm*.

³⁷⁷ Zobacz: <https://raw.githubusercontent.com/twitter/the-algorithm-ml/main/projects/home/recap/README.md#:~:text=contributes%20a%20near,you%20can%20run%20the%20model>.

DZIAŁANIE	WAGA	OPIS
ZGŁOSZENIE	-369,0	Prawdopodobieństwo, że użytkownik_użytkowniczka zgłosi tweeta: znacząco zmniejsza widoczność treści uznanych za obraźliwe lub problematyczne.

Jak wynika z powyższej tabeli, nie wszystkie formy zaangażowania są traktowane jednakowo. Zazwyczaj platforma bardziej ceni interakcje aktywne (np. komentarze lub długi czas przebywania na stronie) niż pasywne (takie jak „polubienie”). Jeśli na przykład model uzna, że użytkownik_użytkowniczka z dużym prawdopodobieństwem odpowie na konkretny tweet, tweet ten zostanie wyżej sklasyfikowany w jego_jej feedzie – komentarz jest postrzegany jako silny wskaźnik zaangażowania. Z kolei jeśli model wykryje wysokie prawdopodobieństwo, że użytkownik_użytkowniczka wyrazi negatywną opinię na temat tweeta, na przykład wyciszając autora_autorkę lub zgłaszając wpis, treść ta znajdzie się niżej w rankingu lub zostanie odfiltrowana³⁷⁸. Należy pamiętać, że ranking dotyczy konkretnego użytkownika_konkretniej użytkowniczki i jego_jej spersonalizowanego kanału informacyjnego, co znaczy, że obniżenie pozycji w rankingu na podstawie przewidywanej negatywnej opinii nie ma charakteru uniwersalnego i nie stanowi odpowiedniego środka naprawczego, który przeciwdziałałby szkodliwym lub nienawistnym treściom na platformie, zamiast je wzmacniać.

Ranking oparty na uczeniu maszynowym dostosowuje oś czasu do każdego użytkownika_każdej użytkowniczki. Dwoje użytkowników z identyczną pulą wytypowanych tweetów wygeneruje różne rankingi, jeśli ich dotychczasowe zachowania różnią się od siebie: model nauczy się różnych profili preferencji dla każdego z nich. Co ważne, model rankingowy podlega okresowym szkoleniom i aktualizacjom (a także ewentualnym modyfikacjom online), aby dostosować się do zmieniających się trendów i gustów użytkowników_użytkowniczek. Na blogu X zaznaczono, że model jest stale udoskonalany na podstawie nowych danych dotyczących interakcji, aby rekomendacje były aktualne i odzwierciedlały „to, co się właśnie dzieje” na platformie³⁷⁹.

Podsumowując, przykładowe wagi wskazują na to, że głównym priorytetem rankingu jest pobudzenie intensywności rozmów i interakcje: oba działania są silnie premiowane. Reakcje negatywne podlegają za to surowym karom. Najwyżej oceniona akcja to „Odpowiedź autora”, która przewyższa resztę stawki z wynikiem +75. Widać więc, że model promuje tweety, które wywołują reakcję autora_autorki.

Po sporządzeniu rankingu tweetów Home Mixer stosuje szereg reguł heurystycznych i biznesowych w celu selekcji i ściślejszego sprofilowania treści wyświetlanych każdemu użytkownikowi_każdej użytkowniczce. Chodzi tu o zapewnienie wystarczającej różnorodności w feedzie każdego użytkownika_każdej użytkowniczki lub usunięcie treści, które naruszają regulamin dotyczący treści bądź politykę X. Na przykład filtry widoczności i bezpieczeństwa wycinają tweety z kont zablokowanych lub wyciszonych przez użytkownika_użytkowniczkę. Inny filtr działa w myśl zasady „opinie sygnalizują zmęczenie treścią”; obniża ocenę niektórych tweetów, jeśli odbiorca_odbiorczyni wyraził_wyraziła na ich temat negatywną opinię, na przykład klikając „Pokaż mniej”³⁸⁰.

7.3 RYZYKO ZWIĄZANE Z ALGORYTMAMI BAZUJĄCYMI NA ZAANGAŻOWANIU

Jak wyjaśniono powyżej, system rekomendacji X zaprojektowano z myślą o maksymalizacji zaangażowania użytkowników_użytkowniczek, mierzonego na podstawie działań takich jak polubienia, retweety, odpowiedzi i czas spędzony na platformie. Zastosowano co prawda pewne środki ograniczające w postaci „kompleksowej heurystyki” (np. filtry bezpieczeństwa), ale te drobne interwencje wiążą się z kompromisami technicznymi i mają drugorzędne znaczenie w stosunku do głównego celu systemu rekomendacji, jakim jest maksymalizacja zaangażowania użytkowników_użytkowniczek. W rezultacie zdolność takich narzędzi do zmniejszenia ryzyka naruszeń praw człowieka jest ograniczona przez nadrzędny cel modelu biznesowego, czyli zwiększanie zaangażowania.

³⁷⁸ Stacey McLachlan, *The X (Twitter) algorithm explained: 2024 guide*, 7 października 2024, <https://blog.hootsuite.com/twitter-algorithm/>.

³⁷⁹ X, *Twitter's recommendation algorithm*.

³⁸⁰ Tamże.

Z powodu modelu biznesowego priorytetowo traktującego zaangażowanie algorytm zaprojektowano w taki sposób, aby zwiększać widoczność treści, które będą wywoływać interakcje. Nawet przy zabezpieczeniach – a wiele z nich zostało ostatnio usuniętych lub znacznie ograniczonych – taki model biznesowy niesie ze sobą poważne zagrożenia dla praw człowieka. Najgroźniejszym aspektem tego modelu jest fakt, że system rekomendacji, mający na celu utrzymanie cyklu zaangażowania, może wzmacniać szkodliwe treści, które wywołują silne reakcje³⁸¹. Większość badań dotyczących amplifikacji algorytmicznej na platformach społecznościowych dowiodła, że jeśli użytkownicy_użytkowniczkowie zaczną wchodzić w interakcje ze szkodliwymi treściami, algorytmy rekomendujące będą im wyświetlać więcej takich treści³⁸².

Przykładowo śledztwo dziennikarskie Washington Post z 2023 roku ujawniło, że konta obserwujące ekstremistów były narażone na kontakt z szeregiem innych rasistowskich i podżegających do przemocy wypowiedzi³⁸³. Wiele użytkowników_użytkowniczek, których wpisy były promowane w sekcji „Dla Ciebie”, zostało zawieszonych przez serwis X, a następnie przywróconych po przejęciu serwisu przez Elona Muska. Nowy szef X zadeklarował ukroczenie szerzenia na platformie mowy nienawiści, twierdząc że: „Nowa polityka Twittera opiera się na wolności słowa, ale nie na wolności zasięgów. Nie znajdziesz danego tweeta, chyba że celowo go szukasz – czyli serwis działa tak, jak reszta internetu”³⁸⁴. Zdaniem Jakuba Szymika skupienie się X na zaangażowaniu ma negatywny wpływ na użytkowników_użytkowniczki platformy:

„Architektura Twittera, oparta na krótkich, zwiezłych komentarzach i polaryzującym algorytmie, wpływa na to, jak ludzie komunikują się w sieci i poza nią. Takie działania mają realne skutki w świecie rzeczywistym. Widzę pomiędzy nimi bardzo głęboką zależność, powinniśmy wziąć to pod uwagę, gdy myślimy o każdej z platform”³⁸⁵.

X uznaje rolę, jaką wzmacnianie odgrywa w jego rekomendacjach: „Rekomendacje mogą wzmacniać treści, dlatego ważne jest, aby były wyświetlane w sposób odpowiedzialny”³⁸⁶. Firma zastrzega również, że wspieranie wartościowych dyskusji jest jedną z podstawowych zasad X: „wolność słowa jest podstawowym prawem człowieka, w przeciwnieństwie do wolności do wzmacniania wypowiedzi w serwisie X”³⁸⁷. Jednakże treści, które nie mogą być rekomendowane (a zatem i wzmacniane) ze względu na zasady platformy, będą nadal dostępne na X dla osób obserwujących autora_autorkę postu oraz w profilu autora_autorki postu³⁸⁸. Treści niekwalifikujące się do rekomendacji to wpisy, które naruszają którąkolwiek z zasad X i które wyjątkowo nie zostały usunięte z platformy ze względu na interes publiczny: może tu chodzić o treści uznane za tylko nieznacznie obraźliwe, szkodliwe lub wprowadzające w błąd³⁸⁹. Konta użytkowników_użytkowniczek mogą zostać objęte zakazem rekomendacji z takich samych powodów³⁹⁰.

X przyznaje użytkownikom_użytkowniczkom ograniczoną kontrolę nad rekomendacjami wyświetlanymi na osiach czasu „Dla Ciebie” i „Obserwowani”. Użytkownicy_użytkowniczki mogą wyciszyć i zablokować powiadomienia na osi „Strona główna” lub zgłosić, że nie są zainteresowani danym postem lub tematem³⁹¹.

Chociaż firma X deklaruje transparentność algorytmu rekomendacji, a w 2023 opublikowała jego kod, niezależny audyt oceny ryzyka przeprowadzony zgodnie z wymogami DSA wykazał, że warunki świadczenia usług przez firmę nie zawierają dostatecznych informacji na temat głównych parametrów stosowanych

³⁸¹ Faculty of Public Health, Response to „Social media, misinformation and harmful algorithms”, inquiry call for evidence, <https://www.fph.org.uk/media/hoejpp0s/social-media-consultation-fph-response.pdf>; Joe Whittaker i in., *What are the links between social media algorithms, generative AI and the spread of harmful content online?*, Written evidence to the UK Parliament Science, Innovation and Technology Committee (SMH0018), 17 grudnia 2024, <https://committees.parliament.uk/writtenevidence/132875/pdf/>.

³⁸² Institute for Strategic Dialogue, *ISD written evidence to the Science, Innovation and Technology Inquiry on Social Media, Misinformation and Harmful Algorithms*, 2025, <https://www.isdglobal.org/wp-content/uploads/2025/01/ISD-Written-Evidence-to-the-Science-Innovation-and-Technology-Committee-Inquiry-on-Social-Media-Misinformation-and-Harmful-Algorithms.pdf>; Joe Whittaker i in., *What are the links between social media algorithms, generative AI and the spread of harmful content online?*.

³⁸³ Washington Post, *Twitter pushes hate speech, extremist content into „For You” pages*.

³⁸⁴ Tamże.

³⁸⁵ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024.

³⁸⁶ X, *About our approach to recommendations*, <https://help.x.com/en/rules-and-policies/recommendations#:~:text=We%20recommend%20posts%20to%20you,by%20those%20in%20your%20network>.

³⁸⁷ X, *About our approach to recommendations*.

³⁸⁸ Tamże.

³⁸⁹ Tamże.

³⁹⁰ Tamże.

³⁹¹ Tamże.

w systemach rekomendacji ani nie objaśniają ich działania. Ich część jest dostępna na stronach poświęconych regułom i polityce firmy, ale nie są to informacje dostatecznie wyczerpujące³⁹². Autorzy audytu zalecili, aby firma X zamieściła w swoich warunkach korzystania z usługi jasne i zrozumiałe objaśnienia parametrów stosowanych w systemie rekomendacji, a także szczegółowe informacje na temat przyjmowanych kryteriów i względnego znaczenia każdego parametru³⁹³.

7.4 KABINY POGŁOSOWE

Z części prac naukowych na temat X wynika, że sposób rekomendowania treści przez tę platformę może sprzyjać tworzeniu się „kabin pogłosowych”³⁹⁴ lub „baniek filtrujących” (informacyjnych), które podsuwają użytkownikom_użytkowniczkom jednolite ideologicznie treści, zazwyczaj zgodne z ich dotychczasowymi przekonaniami³⁹⁵. Efekt kabiny pogłosowej zaobserwowano na wielu platformach społecznościowych, nie jest on charakterystyczny wyłącznie dla X³⁹⁶.

Kluczową cechą takich „kabin” są interakcje między dwójkami użytkowników_użytkowniczek o zbliżonych poglądach: dzięki nim poziom zaangażowania jest wysoki³⁹⁷. Użytkowników_użytkowniczki kabin pogłosowych można zdefiniować jako osoby, które uczestniczą we wspólnym dyskursie, korzystają z tych samych źródeł informacji i mają podobne opinie, a do tego często retweetują swoje wpisy³⁹⁸.

Rozmówcy i rozmówczynie Amnesty International miewali wrażenie, że X tworzy kabiny pogłosowe. Wspomina o tym między innymi Aleksy:

„Zastanawiam się, czy na Twitterze nie jest tak, że gdy widzisz jakiś tweet, algorytm proponuje ci podobne konta. I wpadasz w dziurę bez dna”³⁹⁹.

Studium przypadku z 2024 roku dotyczące efektu kabiny pogłosowej na platformie X wykazało, że użytkownicy_użytkowniczki w kabinach pogłosowych, mimo że stanowią zdecydowaną mniejszość, w znacznym stopniu wpływają na przebieg debat, często rozpowszechniając dezinformację⁴⁰⁰.

Z badania wynika, że skutki tego zjawiska mogą być długotrwałe. Po dwóch latach użytkownicy i użytkowniczki przebywający w obserwowanych przez badaczy kabinach pogłosowych mieli już identyczne opinie oraz znacznie się radykalizowali⁴⁰¹. Autorzy studium zauważyli, że skrajne poglądy badanych nie ograniczały się tylko do pierwotnego tematu, który śledzili (teorie spiskowe dotyczące szczepień przeciwko COVID-19) – po dwóch latach użytkownicy i użytkowniczki mieli skrajne poglądy także w przypadku innych aktualnie kontrowersyjnych tematów, takich jak wojna w Ukrainie, migranci czy kwestie LGBT⁴⁰².

Aby zrozumieć, w jakim stopniu konta marionetkowe w badaniu ilościowym Amnesty International podlegały personalizacji, naukowcy przeanalizowali rodzaj kont rekomendowanych każdej podgrupie w sekcji „Może Ci się spodobać”, a następnie upolitycznienie kont z ich osi „Dla Ciebie”.

Rysunek 7. przedstawia przynależność polityczną kont, które zostały polecane każdej podgrupie kont marionetkowych. Aby zinterpretować tabelę, należy zauważyć, że w przypadku wszystkich kont marionetkowych z grupy „Związki partnerskie” 527 z kont, które im rekomendowano, należało do

³⁹² FTI Consulting, *X Independent Audit*.

³⁹³ Tamże.

³⁹⁴ W kontekście mediów społecznościowych efekt kabiny pogłosowej lub bańka informacyjna to zjawisko, w którym grupa użytkowników_użytkowniczek wchodzi w interakcje głównie z osobami o podobnych przekonaniach, opiniach i poglądach oraz czerpie informacje przede wszystkim od nich. Może to prowadzić do utrwalania istniejących przekonań i ograniczenia kontaktu z różnorodnymi punktami widzenia.

³⁹⁵ Kayla Duskin i in., *Echo chambers in the age of algorithms: an audit of Twitter's friend recommender system*; Manuel Pratelli i in., *Entropy-based detection of Twitter echo chambers*, maj 2024, *PNAS Nexus*, Volume 3, Issue 5, <https://academic.oup.com/pnasnexus/article/3/5/pgae177/7658380>.

³⁹⁶ Manuel Pratelli i in., *Entropy-based detection of Twitter echo chambers*.

³⁹⁷ Tamże.

³⁹⁸ Tamże.

³⁹⁹ Wywiad Amnesty International z Aleksym (pseudonim), 28 lipca 2024.

⁴⁰⁰ Manuel Pratelli i in., *Entropy-based detection of Twitter echo chambers*.

⁴⁰¹ Tamże.

⁴⁰² Tamże, s. 5.

polityków_polityczek popierających wprowadzenie związków partnerskich lub było powiązanych z partiami politycznymi, które się za nim opowiadają (patrz rozdział: „Metodologia”).

Jak ukazano poniżej, dowody personalizacji w rekomendacjach „Może Ci się spodobać” są wyraźne, a rekomendacje jednoznacznie pokrywają się z orientacją polityczną kont marionetkowych. Na tym przykładzie widać, jak łatwo można stworzyć kabinę pogłosową, w której rekomendacje „Może Ci się spodobać” są ściśle powiązane z faktyczną listą kont obserwowanych przez danych użytkowników.

 **RYŚ. 7 PRZYNALEŻNOŚĆ POLITYCZNA KONT PRZEDSTAWIONYCH W REKOMENDACJACH „MOŻE CI SIĘ SPODOBAĆ”**

Grupa	Liczba kont wyświetlanych w rekomendacjach „Może Ci się spodobać”				
	Związki partnerskie	Pełne prawa	Ograniczenie praw	Niejednoznaczne	Nieznane
Związki partnerskie	527	38	4	–	462
Pełne prawa	26	712	1	5	297
Ograniczenie praw	2	735	–	–	277
Niejednoznaczne	48	2	6	350	624

Aby ocenić ryzyko powstania kabin pogłosowych w kanałach „Dla Ciebie”, badacze Amnesty International przeanalizowali stroniczość kont wyświetlonych przez algorytm na osiach czasu każdej grupy. Rysunek 8. przedstawia wyniki tej analizy. Sugeruje ona, że poza tweetami zamieszczonymi bezpośrednio przez Elona Muska istnieją dalsze dowody na personalizację w kanałach „Dla Ciebie” kont marionetkowych. Amnesty International miała trudności z określeniem stroniczości politycznej i poziomu poparcia dla praw osób LGBTI w przypadku wszystkich kont wyświetlanych w kanale „Dla Ciebie” na kontach marionetkowych. Jednak w przypadku kont, które dało się sklasyfikować, istniała wyraźna zgodność między kontami rekomendowanymi użytkownikom_użytkowniczkom marionetkowym za pośrednictwem kanału „Dla Ciebie” a poglądami polityków_polityczek, których obserwowały te konta.

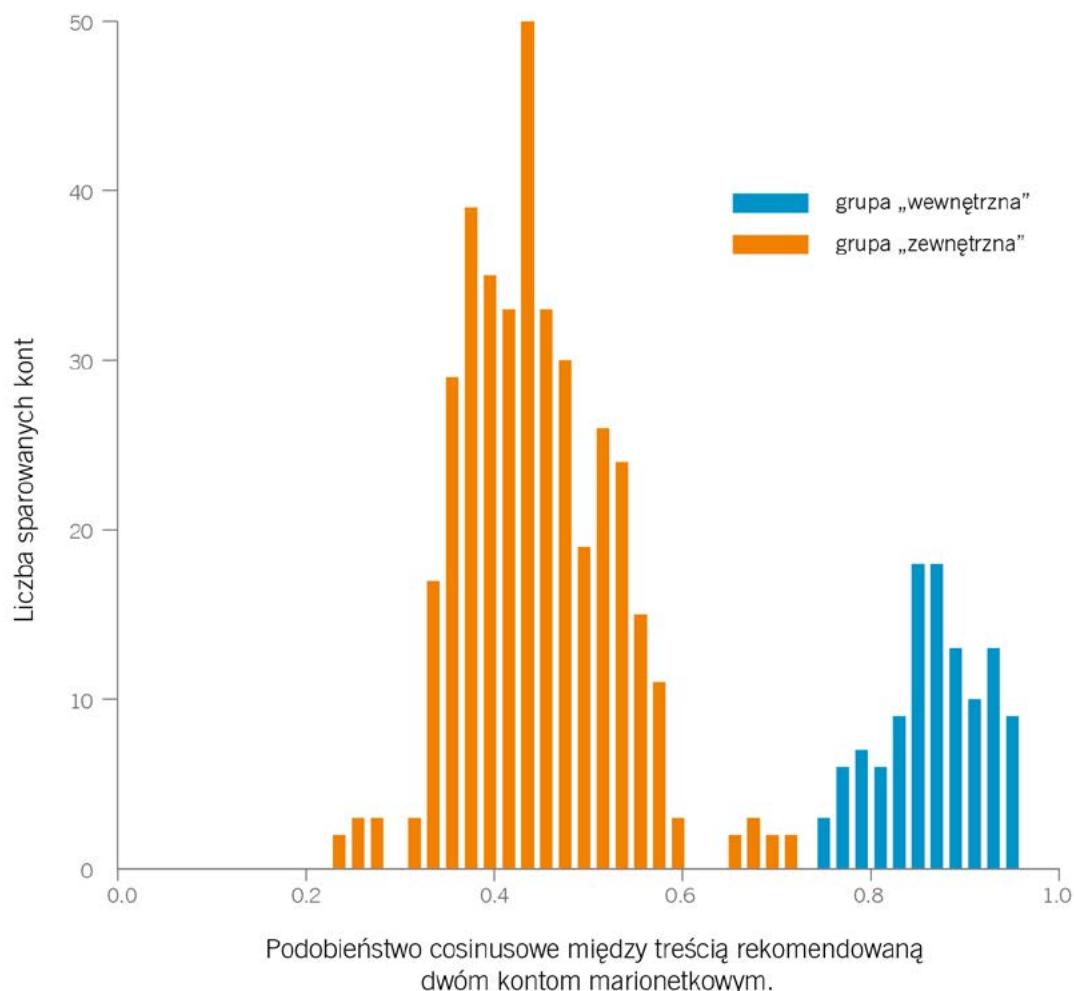
👁️ ↓ RYS. 8



Ponadto Amnesty International przeanalizowała podobieństwo treści wyświetlanych kontom marionetkowym w ramach tej samej grupy, porównując je z treściami kont z pozostałych grup. Badacze porównali treści rekomendowane parom kont marionetkowych (np. konto „Związki partnerskie” nr 1 z kontem „Pełne prawa” nr 1).

Rysunek 9. pokazuje, że kontom marionetkowym, które śledziły ten sam zestaw kont (niebieskie słupki), podsuwane było więcej podobnych treści niż kontom, które tego nie robiły (pomarańczowe słupki). Choć nie komentujemy wprost charakteru tych treści, wynik potwierdza, że rekomendowane treści podlegają faktycznej personalizacji w oparciu o konta, które śledzą konta marionetkowe. Nie są to więc treści wybierane losowo.

👁️ ↓ RYS. 9



7.5 PRIORYTETOWE TRAKTOWANIE IDEI „CYFROWEJ AGORY” KOSZTEM ŚRODKÓW ŁAGODZĄCYCH

OSTRZEŻENIE DOTYCZĄCE TREŚCI

Niniejszy podrozdział zawiera przykłady drastycznych treści nawołujących do przemocy i dyskryminacji, które mogą wywołać silny stres u niektórych czytelników_czytelniczek.

Strategie ograniczania ryzyka stosowane przez X wydają się bazować na chęci utrzymania pozycji platformy jako „cyfrowej agory”; w tym celu serwis stawia na nieograniczoną wolność wypowiedzi w sposób ewidentnie niezgodny z międzynarodowymi normami i prawami człowieka. W pierwszej ocenie ryzyka sporządzonej na podstawie zaleceń DSA platforma jasno dała do zrozumienia, że strategia ta pozostaje kluczowym priorytetem w jej procesach decyzyjnych, informując: „X dąży do tego, aby być placem miejskim internetu

poprzez promowanie i ochronę wolności wypowiedzi. Od początku wiedzieliśmy, że aby osiągnąć ten cel, musimy dać wszystkim możliwość absolutnie swobodnego tworzenia treści i natychmiastowego dzielenia się pomysłami i informacjami”⁴⁰³. W drugiej ocenie ryzyka zgodnej z zaleceniami DSA opisano sposób, w jaki X wybiera środki ograniczające ryzyko ze szczególnym uwzględnieniem ich wpływu na wolność wypowiedzi⁴⁰⁴.

Absolutystyczne podejście X do wolności wypowiedzi jest sprzeczne z międzynarodowym prawem i normami w zakresie praw człowieka. Choć prawo do wolności wypowiedzi musi być chronione, nie jest to prawo absolutne. Musi istnieć równowaga między nim a innymi prawami, takimi jak prawo do niedyskryminacji i prawo do życia wolnego od przemocy ze względu na płeć. Decyzja X o nadaniu priorytetu wolności wypowiedzi, niemal nieskrępowanej jakimikolwiek ograniczeniami, tworzy niedopuszczalny poziom zagrożenia dla użytkowników platformy pochodzących ze społeczności marginalizowanych, w tym polskiej społeczności LGBTI.

Błędne przedkładanie wolności wypowiedzi nad inne prawa człowieka spowodowało, że X podchodzi do moderowania treści w sposób wykraczający poza to, co sam nazywa „binarnym, absolutystycznym podejściem polegającym na usuwaniu/pozostawianiu treści”, a wiele jego strategii łagodzenia skutków, jakie mogą mieć treści szkodliwe, skupia się na ograniczaniu zasięgu treści naruszających zasady platformy⁴⁰⁵. Według X posty objęte ograniczeniami osiągają średnio o 81% mniejszy zasięg lub liczbę wyświetleń niż posty nimi nieobjęte. Platforma stara się również zapobiegać wyświetlaniu reklam obok treści oznaczonych jako szkodliwe⁴⁰⁶.

W ocenie ryzyka z 2024 roku platforma przyznała: „Istnieje ryzyko, że ujawnienie treści prywatnych może mieć wpływ na bezpieczeństwo fizyczne, dobrostan emocjonalny, zdrowie psychiczne i bezpieczeństwo finansowe osób fizycznych”⁴⁰⁷. W niezależnym audycie oceny ryzyka przeprowadzonej przez X, który został przedłożony Komisji Europejskiej w ramach obowiązków spółki wynikających z DSA stwierdzono, że procedura oceny nie była wystarczająco rygorystyczna⁴⁰⁸. Audyt wykazał, że X musi przeprowadzić pełną ocenę ryzyka dla każdego ze swoich systemów rekomendacji w celu zidentyfikowania ryzyka systemowego, zdefiniowania roli i celu działania tychże systemów, ustalenia wskaźników skuteczności oraz ciągłego monitorowania ryzyka stwarzanego przez te systemy⁴⁰⁹. W audycie zalecono również, aby X przeprowadził ocenę ryzyka swojego systemu Freedom of Speech, Not Reach [wolność słowa, a nie zasięgi]⁴¹⁰. Wskazano też, że środki ograniczające ryzyko stosowane przez X nie sprawdzają się przy zmniejszaniu ryzyka systemowego; ponadto brakuje narzędzi łagodzących zagrożenia związane między innymi z funkcjonowaniem systemów algorytmicznych⁴¹¹.

Nieodpowiedzialne i nonszalanckie podejście X do szkodliwych treści potwierdza raport opublikowany w 2024 roku przez polskie Stowarzyszenie „Nigdy Więcej”. Ta organizacja społeczeństwa obywatelskiego jest zarejestrowana jako Trusted Flagger (zaufany zgłaszający) w ramach projektu monitorowania internetu wspieranego finansowo przez program UE Obywatele, Równość, Prawa i Wartości. W okresie od sierpnia 2023 roku do sierpnia 2024 roku „Nigdy Więcej” zgłosiła serwisowi X 343 przypadki treści nawołujących do nienawiści⁴¹². Organizacja początkowo zgłaszała posty za pośrednictwem standardowego interfejsu użytkownika X, a jeśli nie otrzymywała odpowiedzi lub jeśli treści nie były usuwane, przekazywała zgłoszenia za pośrednictwem konta Stowarzyszenia „Nigdy Więcej” na platformie X⁴¹³. W większości przypadków serwis X odmówił usunięcia postów (Stowarzyszenie informuje, że usunięto jedynie 10% zgłoszonych treści)

⁴⁰³ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023, <https://transparency.x.com/content/dam/transparency-twitter/dsa/dsa-sra/dsa-sra-2023/TIUC-DSA-SRA-Report-2023.pdf>.

⁴⁰⁴ Tamże.

⁴⁰⁵ Tamże, s. 8.

⁴⁰⁶ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴⁰⁷ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

⁴⁰⁸ FTI Consulting, *X Independent Audit*.

⁴⁰⁹ Tamże.

⁴¹⁰ Tamże.

⁴¹¹ Tamże.

⁴¹² Stowarzyszenie „Nigdy Więcej”, *The Twitter standards of hate* (sierpień 2023 – sierpień 2024), 2 września 2024, https://www.nigdywiecej.org/docstation/com_docstation/172/the_twitter_standards_of_hate.pdf.

⁴¹³ Tamże.

lub zignorował zgłoszenia⁴¹⁴. Posty – zawierające tekst, obrazy i filmy – można uznać za podżegające do nienawiści wobec mniejszości, w tym osób LGBTI⁴¹⁵.

Niektóre posty zgłoszone przez Stowarzyszenie „Nigdy Więcej” do serwisu X i mimo zgłoszenia nieusunięte były skierowane konkretnie przeciwko społeczności LGBTI. Można je uznać za podżeganie do przemocy i nawoływanie do nienawiści. Na przykład jeden z zgłoszonych wpisów, w stosunku do którego serwis X nie podjął żadnych działań, brzmiał: „Jebać gender. Jebać kurwy zboczone. Jebać transwestytów. Do pieca kurwy !!!!!”⁴¹⁶.

Kilka postów zgłoszonych przez Stowarzyszenie w okresie od sierpnia 2023 do sierpnia 2024 roku nadal, czyli do maja 2025 roku, pozostaje widocznych na platformie. Tweety te zostały udokumentowane poniżej; przedstawiają społeczność LGBTI jako zboczeńców, zawierają obelgi i nawołują do dyskryminacji społeczności LGBTI.



Post Antoniego Kocemby⁴¹⁷



Post Konfederacji⁴¹⁸

⁴¹⁴ Tamże.

⁴¹⁵ Tamże.

⁴¹⁶ Stowarzyszenie „Nigdy Więcej”, The Twitter standards of hate, sierpień 2023 – sierpień 2024.

⁴¹⁷ Antoni Kocemba, post na platformie X, 31 lipca 2023, https://x.com/antoni_kocemba/status/1685950832657797120.

⁴¹⁸ Konfederacja, post na platformie X, 28 lipca 2023, https://x.com/KONFEDERACJA_/status/1684882568087543808.



Post arcziwyspa ⁴¹⁹



Post Seliana ⁴²⁰

Powyższe posty, w maju 2025 roku nadal krążące na platformie, są wyraźnym dowodem występowania na X szkodliwych treści, które stały się tam normą ze względu na bezkrytyczne podejście firmy do wolności słowa. X wykorzystuje wolność słowa jako argument, aby usprawiedliwić swoją niedbałość w moderowaniu treści. Nawet w przypadku otrzymywania zgłoszeń dotyczących wpisów, które można uznać za podżeganie do przemocy i nawiązanie do nienawiści wobec społeczności LGBTI, X wydaje się ignorować powszechność szkodliwych treści na platformie, nie biorąc pod uwagę ryzyka, jakie treści te stanowią dla praw osób

⁴¹⁹ arcziwyspa, post na platformie X, 17 października 2023, <https://x.com/ArcziWyspa/status/1714198139186385141>.

⁴²⁰ Selian, post na platformie X, 3 grudnia 2023, <https://x.com/Selianski/status/1731452869792924087>.

marginalizowanych. Dotyczy to również ich prawa do wolności wypowiedzi – rozmówcy i rozmówczynie Amnesty International ze społeczności LGBTI wspominali o tym, że stosują na platformie X autocenzurę. Brak należytego uwzględnienia praw innych niż wolność wypowiedzi znajduje odzwierciedlenie w ocenach ryzyka X w latach 2023 i 2024, które zgodnie z międzynarodowymi normami nie spełniają wymogów należytej staranności w zakresie poszanowania praw człowieka.

7.6 BRAK KONTAKTÓW X ZE SPOŁECZEŃSTWEM OBYWATELSKIM

Ważnym czynnikiem przy ocenie odpowiedzialności X za naruszanie praw człowieka społeczności LGBTI w Polsce jest przewidywanie udziału przedsiębiorstwa w naruszaniu tych praw. Zgodnie z międzynarodowymi normami, jeśli przedsiębiorstwo wie, a powinno wiedzieć, że istnieje ryzyko, iż przyczyni się do naruszenia praw człowieka, ma ono obowiązek podjąć niezbędne kroki w celu zapobieżenia temu naruszeniu oraz wykorzystać swoje wpływy, aby w możliwie największym stopniu złagodzić jej wszelkie negatywne skutki⁴²¹. Dlatego zachęca się przedsiębiorstwa do współpracy z zainteresowanymi stronami w celu identyfikacji i ograniczenia ryzyka. Zaangażowanie zainteresowanych stron jest również niezbędnym elementem przeprowadzania oceny ryzyka na mocy DSA.

Tymczasem Amnesty International ustaliła, że co najmniej od 2022 roku X wykazuje bardzo niewielkie zaangażowanie w aktywny dialog z polskimi organizacjami społeczeństwa obywatelskiego działającymi na rzecz społeczności LGBTI w celu omawiania sposobów minimalizacji zagrożeń na platformie. Na przykład osoba pracująca w jednym z najbardziej znanych polskich NGO zajmujących się kwestiami LGBTI powiedziała Amnesty International, że nie słyszała o jakiegokolwiek komunikacji między tą organizacją a X⁴²². Również Mateusz Kaczmarek, członek zarządu Grupy Stonewall, powiedział Amnesty, że X ani razu nie skontaktował się z Grupą w celu omówienia potencjalnych zagrożeń lub środków ograniczających ryzyko⁴²³.

Julia Kata powiedziała Amnesty International, że nie ma żadnych informacji o konsultacjach między X a organizacjami społecznymi LGBTI w Polsce:

„My [polskie organizacje LGBTI] siedzimy w tym razem, komunikujemy się ze sobą i prawdopodobnie gdyby X skontaktował się z jedną, dwiema lub trzema organizacjami, pozostałe by się o tym dowiedziały. I poprosiłyby o przekazanie danych kontaktowych [do osób z X], bo chcielibyśmy z nim porozmawiać”⁴²⁴.

Niewielkie zainteresowanie X polskim społeczeństwem obywatelskim znajduje odzwierciedlenie w jego ocenie ryzyka na mocy DSA na rok 2024, w której poinformowano, że firma weszła w nieliczne interakcje z organizacjami społeczeństwa obywatelskiego; nie podano jednak szczegółowych informacji na temat liczby takich działań ani obszarów wiedzy specjalistycznej lub konkretnych społeczności, których dotyczyły⁴²⁵. Ponadto opis kontaktów X z organizacjami społeczeństwa obywatelskiego wydaje się skupiać na działaniach mających na celu przeszkolenie tych organizacji w zakresie lepszego wykorzystywania narzędzi zgłaszania treści na platformie, a nie na wykorzystaniu specjalistycznej wiedzy organizacji na temat szkodliwych treści i marginalizowanych społeczności⁴²⁶.

7.7 NIEWYSTARCZAJĄCE ŁAGODZENIE RYZYKA SYSTEMOWEGO

W ocenie ryzyka na mocy DSA na rok 2024 firma X poinformowała, że stosowane przez nią środki kontroli ograniczają poziom ryzyka w większości zidentyfikowanych obszarów do niskiego lub średniego⁴²⁷. Jednak obecne i planowane środki ograniczające ryzyko przedstawione w ocenie dotyczą jedynie ulepszeń polityk, systemów moderacji treści (w tym wdrażania i wykrywania) oraz działań podnoszących świadomość na temat „Uwag społeczności”⁴²⁸. Środki te w niedostatecznym stopniu uwzględniają zagrożenia związane z modelem biznesowym X, w tym nacisk na algorytmiczną optymalizację zaangażowania czy ryzyko, które

⁴²¹ Wytyczne dotyczące biznesu i praw człowieka, zasada 19. i komentarz.

⁴²² Wywiad Amnesty International z Aleksym (pseudonim), 26 lipca 2024.

⁴²³ Wywiad Amnesty International z Mateuszem Kaczmakiem, 28 lipca 2024.

⁴²⁴ Wywiad Amnesty International z Julią Katą, 29 lipca 2024.

⁴²⁵ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

⁴²⁶ Tamże.

⁴²⁷ Tamże.

⁴²⁸ Tamże.

obecna działalność firmy stwarza dla społeczności marginalizowanych, na przykład poprzez przeznaczanie niewystarczających zasobów na moderację treści.

X twierdzi, że jego systemy rekomendacji zaprojektowano tak, aby wykluczać treści szkodliwe i naruszające zasady poprzez integrację z systemami filtrowania widoczności i innymi rozwiązaniami. Wykorzystuje się przy tym modele przewidywania jakości treści, co ma zapobiegać uzyskiwaniu przez treści szkodliwe i naruszające zasady wyższych pozycji w rankingach⁴²⁹. W marcu 2023 roku X wdrożył politykę mającą na celu usuwanie z platformy treści zawierających przemoc i mowę nienawiści⁴³⁰. Wydaje się jednak, że jeśli systemy rekomendacji nieprawidłowo dopuszczają algorytmiczne promowanie szkodliwych treści, pozostaje niewiele skutecznych środków łagodzących, które minimalizowałyby szkody. Dzieje się tak, bo zgodnie z własną oceną ryzyka platforma w dużym stopniu bazuje na kontroli przyznawanej użytkownikom_ użytkownikom, sprowadzającej się do możliwości wyciszania powiadomień lub ograniczania odpowiedzi na posty⁴³¹.

Poleganie na ulepszeniach polityki – szczególnie biorąc pod uwagę, że coraz bardziej liberalne podejście do szkodliwych treści doprowadziło do pogorszenia jakości tejże polityki – okazało się nieskuteczne w ograniczaniu ryzyka systemowego na platformie. I tak na przykład pomimo polityki usuwania mowy nienawiści podlegającej do przemocy większość aktywistów LGBTI, z którymi rozmawiała Amnesty International, zgłaszała, że przez kilka lat wielokrotnie spotykała się z takimi wypowiedziami na X lub była ich bezpośrednim celem.

X dostrzega ryzyko, że „personalizacja rekomendowanych treści może w niektórych okolicznościach przyczyniać się do powstawania baniek informacyjnych, ograniczając dostęp użytkowników do pluralistycznych źródeł informacji”⁴³², ale nie przedstawia żadnych konkretnych środków zaradczych.

Ponadto X zauważa, że komentarze, podobnie jak posty, mogą stanowić zagrożenie dla użytkowników_ użytkowniczek platformy celowo wystawianych na kontakt z komentarzami zawierającymi mowę nienawiści: oznaczenie autora_ autorki oryginalnego posta powoduje bowiem jego_ jej powiadomienie⁴³³. Co więcej, zgodnie z najnowszą oceną ryzyka X, platforma uznaje mowę nienawiści za „treści niezgodne z prawem” w rozumieniu przepisów DSA⁴³⁴. Jednakże z racji tego, że Polska nie uwzględnia w przepisach penalizujących mowę nienawiści przesłanek dyskryminacyjnych dotyczących osób LGBTI⁴³⁵, nie jest jasne, jak platforma postępowałaby z treściami nienawistnymi wymierzonymi w osoby LGBTI, gdyby nie były one powiązane z nawoływaniem do przemocy.

Kwestia ta budzi dodatkowe obawy, ponieważ X w dużym stopniu polega na automatycznym wykrywaniu naruszeń swoich zasad⁴³⁶. W przypadku obelg i stereotypów firma korzysta ze słowników opracowanych specjalnie z myślą o językach UE⁴³⁷. Nie jest to jednak zasób wystarczający do moderowania treści, zwłaszcza że X zatrudnia tylko dwóch moderatorów mówiących po polsku⁴³⁸.

X twierdzi, że to dzięki strategiom ograniczającym „[...] 99,99% wyświetleń dotyczy treści uznanych za »zdrowe«. Mniej niż 0,01% wyświetlonych postów zawiera mowę nienawiści”⁴³⁹. Firma nie podaje jednak szczegółowego rozkładu tych danych ze względu na język lub kraj.

Jakub Szymik powiedział Amnesty International, że widział, jakie szkody mogą wyrządzić nienawistne komentarze na X:

⁴²⁹ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴³⁰ X, *Violent Content policy*.

⁴³¹ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴³² X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴³³ Tamże.

⁴³⁴ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

⁴³⁵ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴³⁶ Tamże.

⁴³⁷ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴³⁸ X, *DSA Transparency Report – April 2025*.

⁴³⁹ Tamże.

„Współpracuję z organizacją LGBTQ, której liderem jest osoba bardzo widoczna na platformie i która korzysta z Twittera do nagłaśniania swojej działalności. Widzę, jak ogromny wpływ mają na niego lawiny anonimowych komentarzy, ale także wypowiedzi osób podających swoje prawdziwe dane i publicznie nawołujących do przemocy lub wysyłających groźby. Masowa transmisja takich treści i fala niezwykle agresywnych wypowiedzi bardzo odbijają się na nim i na całej organizacji”⁴⁴⁰.

Szymik dodał, że często widzi na X komentarze skierowane przeciwko aktywistom LGBTI:

„Większość sytuacji, z którymi się spotykam, dotyczy konkretnej osoby, która zabiera głos, po czym zbiera mnóstwo komentarzy o bardzo agresywnym charakterze. Na przykład: »ktoś powinien cię zastrzelić, zabić cię, nie powinienś mieć prawa się wypowiadać«”⁴⁴¹.

Środkiem ograniczającym ryzyko pojawienia się szkodliwych treści w komentarzach jest kontrola odpowiedzi, która pozwala użytkownikowi ograniczyć krąg osób mogących odpowiadać na jego_jej posty – można zezwolić na to wyłącznie osobom wymienionym w poście lub całkowicie wyłączyć możliwość odpowiadania⁴⁴².

Magda Dropek wyjaśniła Amnesty International, że takie narzędzia mają charakter prewencyjny i nie wystarczają do odpowiedniej reakcji na naruszenia:

„W ostatnich latach zauważyłam w swoich socialach pewną prawidłowość – oczywiście bardzo trudno jest jakoś zaradzić wyjątkowo nienawistnym wiadomościom. W moim przypadku, gdy ludzie piszą do mnie setki komentarzy i wiadomości typu »zabij się«, »nikt cię tu nie chce«, »w tym kraju jesteś jak śmieć [...]«, mam narzędzia, aby sobie z takim zalewem poradzić. Ale dla mnie ważne jest to, że bardzo często społeczność, która mnie śledzi, też widzi te wiadomości. Czuję się przez to całkowicie bezbronna, zwłaszcza po tym, jak Twitter zmienił nazwę na X, bo narzędzia [po zmianie nazwy formy] stały się bardzo trudne w obsłudze”⁴⁴³.

W ocenie ryzyka na 2024 rok X informuje, że może to wywołać u użytkowników poczucie strachu i zastraszenia oraz prowadzić do autocenzury, a także zauważa, że platforma może być nadużywana do promowania nienawiści lub podżegania do wrogości, dyskryminacji i przemocy⁴⁴⁴, czego doświadczyły osoby LGBTI, z którymi Amnesty International przeprowadziła wywiady.

Warto raz jeszcze powtórzyć, że środki ograniczające takie zagrożenia w X są niewystarczające, bo sprowadzają się do weryfikacji zasad i procesów oraz funkcji „Uwagi społeczności”, która w zasadzie przekazuje moderowanie treści w ręce użytkowników_użytkowniczek X⁴⁴⁵. Jak wspomniano w sekcji 5.4.1, funkcja „Uwagi społeczności” ma poważne ograniczenia i wady.

7.8 ŚWIADOMOŚĆ X W KWESTII ZAGROZEŃ SYSTEMOWYCH

Z najnowszej oceny ryzyka zgodnej z DSA wynika, że firma X jest w pełni świadoma, iż jej platforma stanowi systemowe zagrożenie dla wielu praw człowieka, w tym zagrożenie „na poziomie społecznym”, w szczególności dla społeczności marginalizowanych⁴⁴⁶. X twierdzi, że jego „podejście do oceny i ograniczania ryzyka związanego ze szkodliwymi treściami nadal opiera się na zestawie zasad uwzględniających szkody fizyczne, psychiczne, informacyjne, ekonomiczne i społeczne, co pozwala [...] analizować potencjalne szkody wynikające z treści i zachowań, które mogą wystąpić w serwisie X”⁴⁴⁷.

Chociaż algorytmicznie wzmocnianie i rekomendowanie treści stanowią kluczowe elementy modelu biznesowego X, platforma utrzymuje, że jej algorytmy nie promują celowo treści zawierających „obraźliwe

⁴⁴⁰ Rozmowa wideo Amnesty International z Jakubem Szymikiem, 5 sierpnia 2024 r.

⁴⁴¹ Tamże.

⁴⁴² X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴⁴³ Wywiad Amnesty International z Magdą Dropek, 24 lipca 2024.

⁴⁴⁴ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

⁴⁴⁵ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, sierpień 2024.

⁴⁴⁶ Tamże.

⁴⁴⁷ Tamże.

określenia” i „nienawistne zwroty”⁴⁴⁸. Firma przyznaje jednak, że wcześniejsze badania wykazały, iż „w pewnych okolicznościach [jej] systemy rekomendacji mogą prowadzić do wzmocnienia kont o określonych poglądach ideologicznych kosztem innych. Jednakże mimo potencjalnie wpisanego w te systemy ryzyka stronniczości wspomniane analizy nie wykazały istnienia jednoznacznych, pojedynczych czynników wpływających na ten efekt. W innych okolicznościach ten sam algorytm wywierał odmienny wpływ na treści polityczne”⁴⁴⁹. Jeśli systemy rekomendacji rzeczywiście funkcjonują w różny sposób w rozmaitych kontekstach, świadczy to o konieczności przeprowadzenia przez X analizy zgodności z prawami człowieka w poszczególnych krajach pod kątem potencjalnych szkodliwych skutków stosowania tychże systemów.

X ma świadomość, że do nękania mogą być wykorzystywane różne elementy jego sposobu funkcjonowania, takie jak wzmianki i cytaty, „co stanowi zagrożenie dla godności ludzkiej, braku dyskryminacji oraz poszanowania życia prywatnego i rodzinnego”⁴⁵⁰. X uznaje ponadto, że „cyfrowa przepaść między płciami mogła [...] przyczynić się do tego, że kobiety i członkowie społeczności LGBTQ+ stali się celem nienawiści i nadużyć”⁴⁵¹.

7.9 OCENA PRZYZCYNENIA SIĘ X DO AKTÓW TFGBV PRZECIWKO SPOŁECZNOŚCI LGBTI W POLSCE

Zgodnie z Wytycznymi ONZ przedsiębiorstwo ma negatywny wpływ na prawa człowieka, gdy jego działania (w tym zaniechania) w istotny sposób zwiększają ryzyko wystąpienia określonego skutku – nawet gdy działania przedsiębiorstwa same w sobie nie są wystarczające, aby spowodować takowy skutek⁴⁵². Żeby wypełnić obowiązek poszanowania praw człowieka, X musi stosować się do wytycznych, które wymagają od przedsiębiorstw, aby te „uniknęły powodowania lub przyczyniania się do negatywnego wpływu na prawa człowieka przez swoją działalność” oraz „podejmowały działania mające na celu zapobiegnięcie negatywnemu wpływowi na prawa człowieka powiązanemu bezpośrednio z ich działalnością, produktami lub usługami poprzez ich relacje biznesowe lub łagodzący taki wpływ, nawet jeśli one same nie przyczyniły się do niego”⁴⁵³.

W latach 2019–2023 z serwisu X korzystało w Polsce wiele osób, w tym przedstawiciele rządu, władz regionalnych i aktywiści anti-LGBTI, którzy publikowali treści wymierzone w społeczność LGBTI. Niektóre z tych treści podlegały do przemocy i dyskryminacji. Chociaż po wyborach parlamentarnych w 2023 roku retoryka polityczna wobec społeczności LGBTI stała się bardziej przyjazna, skutki wieloletniej mowy nienawiści są nadal widoczne na platformie, a osoby LGBTI pozostają celem TFGBV.

Udział X w naruszeniach praw człowieka, których doświadczyła polska społeczność LGBTI, wynika z faktu, że środki łagodzące stosowane przez serwis – takie jak moderowanie treści – nie rozwiązały w wystarczającym stopniu problemu powszechnego występowania na platformie TFGBV, w tym gróźb przemocy, nękania online i doxingu.

Skutki są tym bardziej dotkliwe, że X jest w Polsce ważną platformą dyskursu politycznego, a także źródłem informacji dla dziennikarzy i aktywistów. Można uznać, że X miał niekorzystny wpływ na prawa człowieka również ze względu na swoje podejście do przewidywania ryzyka, które stwarzała jego działalność. Pomimo dobrze udokumentowanych ataków na społeczność LGBTI ze strony polityków i polityczek reprezentujących wysokie urzędy w Polsce X nie zdołał odpowiednio ograniczyć ryzyka naruszania praw człowieka w związku ze swoją działalnością w tym kraju.

Istnieje jeszcze wiele kroków, które firma X mogła podjąć, aby zapobiec rozprzestrzenianiu się i rozpowszechnianiu treści skierowanych przeciwko społeczności LGBTI na platformie; należą do nich na przykład bardziej aktywne stosowanie mechanizmów moderacji treści. W sierpniu 2024 roku Amnesty International wysłała do X pismo z prośbą o informacje na temat personelu i zasobów X przeznaczonych

⁴⁴⁸ Tamże.

⁴⁴⁹ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴⁵⁰ Tamże.

⁴⁵¹ X, *Report Setting Out the Results of Twitter International Unlimited Company Risk Assessment Pursuant to Article 34 of the Digital Services Act*, wrzesień 2023.

⁴⁵² Business & Human Rights Resource Centre, *Practical Definitions of Cause, Contribute, and Directly Linked to Inform Business Respect for Human Rights*, 9 lutego 2017, <https://www.business-humanrights.org/en/latest-news/practical-definitions-of-cause-contribute-and-directly-linked-to-inform-business-respect-for-human-rights/>

⁴⁵³ Wytyczne dotyczące biznesu i praw człowieka, zasada 13. i komentarz.

na działalność w Polsce w latach 2019–2024, w tym o liczbę moderatorów treści w Polsce, ich znajomość języka polskiego i fizyczne miejsce pracy, ale firma nie udzieliła odpowiedzi⁴⁵⁴. W niniejszym raporcie opisano ze szczegółami, że X nie był w stanie odpowiednio moderować treści w Polsce. Ponadto platforma powoli reagowała na opinie własnych użytkowników_użytkowniczek i organizacji społeczeństwa obywatelskiego monitorującej mowę nienawiści w internecie. Obie te grupy zgłaszały treści, które powinny być uznane za przykłady TFGBV, ale na niektóre zgłoszenia firma w ogóle nie reagowała. W rezultacie szkodliwe treści mogły dalej krążyć na platformie, a niektórzy członkowie społeczności LGBTI przestali zgłaszać TFGBV z powodu braku reakcji ze strony X.

Pomimo wynikającego z DSA obowiązku identyfikowania ryzyka systemowego niewiele wskazuje na to, aby X podjął istotne działania w kierunku odpowiedniej identyfikacji lub zmniejszenia ryzyka, jakie platforma stwarza dla społeczności LGBTI w Polsce.

Analiza Amnesty International, oparta na międzynarodowych normach praw człowieka, w tym wytycznych ONZ, dotycząca roli X w łamaniu praw człowieka wobec społeczności LGBTI w Polsce od 2019 roku do chwili obecnej, pozwala wyciągnąć następujące wnioski:

1. Z racji tego, że X jest kluczową platformą dla polskich polityków_polityczek, dziennikarzy_dziennikarek i aktywistów_aktywistek, członkowie i członkinie polskiego rządu, partie polityczne oraz aktywiści i aktywistki anty-LGBTI wykorzystali_wykorzystały X do publikowania treści uderzających w społeczność LGBTI. Niektóre z ich wpisów podlegały do przemocy i dyskryminacji.
2. Brak moderacji treści przez X w Polsce pozwolił na rozpowszechnianie na platformie treści podlegających do przemocy i dyskryminacji wobec społeczności LGBTI.
3. X wiedział lub powinien był wiedzieć, że ryzykuje współudziałem w łamaniu praw człowieka w Polsce, zwłaszcza że działania firmy w zakresie moderacji treści w tym kraju są finansowane gorzej niż w innych państwach europejskich.
4. X nie dołożył należytej staranności w przestrzeganiu praw człowieka, która to staranność mogła lub powinna była pozwolić na zidentyfikowanie ryzyka związanego z działalnością firmy w Polsce. X nie wdrożył również odpowiednich środków naprawczych, które mogłyby zapobiec szkodom wyrządzonym w Polsce lub je złagodzić.
5. Z analizy przypadków opisanych w rozdziale 6. wynika, że brak należytej staranności X w kwestii rozpowszechniania treści podlegających do przemocy, dyskryminacji i nienawiści oraz nieadekwatność działań w zakresie moderacji treści przyczyniły się do naruszenia szeregu praw człowieka, w tym prawa do wolności wypowiedzi, prawa do równości i niedyskryminacji oraz prawa do ochrony zdrowia.

X przyczynił się do aktów TFGBV na polską społeczność LGBTI, w związku z czym ponosi odpowiedzialność za zadośćuczynienie za wyrządzoną krzywdę.

⁴⁵⁴ Pismo Amnesty International skierowane do firmy X, 22 sierpnia 2024.

8. ZADOŚĆUCZYNIENIE I UNIKANIE SZKÓD W PRZYSZŁOŚCI

W rozdziale 7. opisaliśmy, w jaki sposób firma X przyczyniła się i nadal przyczynia się do naruszeń praw człowieka związanych z TFGBV, czyli przemocy ze względu na płeć wspomaganej technologicznie, której doświadcza społeczność LGBTI w Polsce. W związku z tym, zgodnie z wytycznymi ONZ, firma X ma obowiązek skutecznie zadośćuczynić osobom, które ucierpiały w wyniku jej działalności. Ponadto zgodnie z wymogami DSA X ma obowiązek zidentyfikować i ograniczyć systemowe zagrożenia dla praw podstawowych, które mogą wynikać z jego działalności.

W niniejszym rozdziale analizujemy, jak X może wypełnić swoje obowiązki wynikające z międzynarodowych norm dotyczących biznesu i praw człowieka oraz z prawa europejskiego.

8.1 ODPOWIEDZIALNOŚĆ X ZA DZIAŁANIA NAPRAWCZE

8.1.1 MIĘDZYNARODOWE NORMY DOTYCZĄCE PRAW CZŁOWIEKA

Przedsiębiorstwa, które przyczyniły się do negatywnego wpływu na prawa człowieka, mają obowiązek zapewnić odpowiednie środki zaradcze podmiotom poszkodowanym⁴⁵⁵. Adekwatny środek zaradczy zależy od rodzaju szkody i może przybierać różne formy, w tym przeprosin, restytucji, rehabilitacji, rekompensaty finansowej lub niefinansowej oraz sankcji (karnych lub administracyjnych, takich jak grzywna), jak również gwarancji niewystąpienia dalszych przypadków łamania prawa⁴⁵⁶.

Publiczne przeprosiny są ważną formą zadośćuczynienia, która polega na uznaniu faktów i przyjęciu za nie odpowiedzialności; może im towarzyszyć weryfikacja faktów oraz pełne i publiczne ujawnienie prawdy⁴⁵⁷.

Równie ważną formą zadośćuczynienia jest gwarancja niedopuszczenia do powtórzenia się naruszeń w przyszłości.

W tym kontekście zapobieganie dalszym nadużyciom można osiągnąć w wyniku zastosowania różnych środków, do których należą: działania regulacyjne i wynikające z przyjęcia odpowiedzialności ze strony władz państwowych oraz działania samych przedsiębiorstw – wszystkie mogą się przyczynić do zagwarantowania niepowtórzenia się nadużyć⁴⁵⁸. Do innych możliwych środków należą przeznaczenie odpowiednich zasobów na działania związane z moderowaniem treści oraz podjęcie konstruktywnego dialogu z organizacjami społeczeństwa obywatelskiego w celu lepszego zrozumienia i rozwiązania problemu TFGBV.

8.1.2 POSTANOWIENIA DSA

Zgodnie z wytycznymi ONZ państwa są zobowiązane do podjęcia „odpowiednich kroków w celu zapobiegania, badania, karania i naprawiania” naruszeń praw człowieka związanych z działalnością

⁴⁵⁵ Wytyczne OECD dotyczące należytej staranności; *Wytyczne dotyczące biznesu i praw człowieka*, zasada 22.

⁴⁵⁶ *Wytyczne dotyczące biznesu i praw człowieka*, zasada 25.

⁴⁵⁷ *UN Basic Principles and Guidelines on the Right to Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law*, 21 marca 2006, UN Doc. A/RES/60/147, zasada 22.

⁴⁵⁸ Amnesty International, *Injustice Incorporated: Corporate Abuses and the Human Right to Remedy*, Index: POL 30/001/2014, 7 marca 2014, <https://www.amnesty.org/en/documents/pol30/001/2014/en/>, s. 18.

gospodarczą na swoim terytorium lub terytorium podlegającym ich jurysdykcji⁴⁵⁹. DSA zapewnia użytkownikom i użytkowniczkom platform określone środki odwoławcze. W przypadku naruszenia przepisów DSA użytkownicy i użytkowniczki i wszelkie organizacje upoważnione do występowania w ich imieniu mają prawo złożyć skargę na przedsiębiorstwo do koordynatora ds. usług cyfrowych (DSC) w kraju zamieszkania⁴⁶⁰. Użytkownicy i użytkowniczki platform mają również prawo do otrzymania odszkodowania od przedsiębiorstw za szkody lub straty wynikające z naruszenia przez nie przepisów DSA⁴⁶¹.

8.2 WYPEŁNIANIE PRZEZ X OBOWIĄZKÓW WYNIKAJĄCYCH Z DSA

Zgodnie z Artykułem 34. DSA bardzo duże platformy internetowe (VLOP), takie jak X, ze względu na swoją skalę i potencjalny wpływ, jaki mogą wywierać na społeczeństwo, mają w UE obowiązek przeprowadzania systemowej oceny ryzyka swoich usług⁴⁶². Artykuł 34. nakłada na serwisy takie jak X obowiązek „starannego identyfikowania, analizowania i oceny wszelkiego ryzyka systemowego [...] wynikającego ze struktury i funkcjonowania ich usług i powiązanych systemów, w tym systemów algorytmicznych, oraz z korzystania z ich usług”⁴⁶³. Zakres ryzyka dotyczy wszelkich „rzeczywistych i przewidywalnych negatywnych skutków dla korzystania z praw podstawowych, takich jak wolność wypowiedzi, wolność mediów i pluralizm, dyskryminacja, ochrona konsumentów i prawa dzieci”⁴⁶⁴. Przeprowadzając ostatnią ocenę ryzyka, X powinien był w szczególności wziąć pod uwagę, czy i w jaki sposób czynniki takie jak konstrukcja jego systemów rekomendacji, wszelkie pozostałe istotne systemy algorytmiczne oraz między innymi systemy moderowania treści wpływają na ryzyko systemowe naruszenia praw podstawowych użytkowników i użytkowniczek, a także na zagrożenie przemocą ze względu na płeć oraz ryzyko zaszkodzenia ich dobrostanowi fizycznemu i psychicznemu.

Tymczasem ocena ryzyka przeprowadzona przez X w 2024 roku nie uwzględniała tych zagrożeń w wystarczającym stopniu. Nie uwzględniono w niej również zagrożeń dla praw społeczności LGBTI wynikających ze stosowania systemów rekomendacji, a także z braku zasobów na moderowanie treści. Zgodnie z artykułem 35. DSA VLOP powinny wdrażać racjonalne, proporcjonalne i skuteczne środki ograniczające ryzyko, dostosowane do konkretnych zagrożeń systemowych określonych w artykule 34. Środki te mogą obejmować: dostosowanie architektury, funkcji lub funkcjonowania usług, w tym interfejsów cyfrowych; dostosowanie procesów moderacji treści; oraz testowanie i dostosowanie systemów algorytmicznych, w tym systemów rekomendacji⁴⁶⁵. Mimo że X nie zidentyfikował w wystarczającym stopniu ryzyka systemowego, jakie stwarza dla społeczności LGBTI w Polsce, jednak podjęcie przezeń wymienionych środków w ramach rzetelnego zadośćuczynienia mogłoby zagwarantować, że powstałe szkody się nie powtórzą.

Kluczowym aspektem przeprowadzania analiz należytej staranności w zakresie praw człowieka, takich jak ocena ryzyka, jest zaangażowanie interesariuszy. W punkcie 90. preambuły DSA podkreślono, że zaangażowanie społeczeństwa obywatelskiego i grup szczególnie dotkniętych problemem jest niezbędne do tego, aby zapewnić, że ocena ryzyka i środki ograniczające zagrożenia będą bazować na „najlepszych dostępnych informacjach”⁴⁶⁶. Wszelkie założenia dotyczące ryzyka systemowego muszą zostać sprawdzone w szczególności z grupami najbardziej narażonymi na takie ryzyko⁴⁶⁷. Tymczasem X nie konsultował oceny ryzyka z żadną z organizacji społeczeństwa obywatelskiego, z którymi Amnesty International rozmawiała w ramach niniejszego badania, a sama ocena wskazuje na nikłe interakcje firmy z tymi organizacjami.

⁴⁵⁹ Amnesty International, *What the EU's Digital Services Act Means for Human Rights*.

⁴⁶⁰ Tamże.

⁴⁶¹ Tamże.

⁴⁶² Akt o usługach cyfrowych, art. 34.; Komisja Europejska, *DSA: Very large online platforms and search engines*, 12 lutego 2025, <https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops#:~:text=Those%20designated%20as%20VLOPs%20or,respond%20to%20the%20auditor%27s%20recommendations>.

⁴⁶³ Akt o usługach cyfrowych, art. 34.

⁴⁶⁴ Tamże; Komisja Europejska, *DSA: Very large online platforms and search engines*.

⁴⁶⁵ *Akt o usługach cyfrowych*, art. 35.

⁴⁶⁶ Tamże, pkt 90. preambuły.

⁴⁶⁷ Tamże, pkt 90. preambuły.

8.3 KARY PRZEWIDZIANE W DSA

Zgodnie z DSA Komisja Europejska posiada bezpośrednie uprawnienia nadzorcze i egzekucyjne, które może zastosować wobec dostawców usług nieprzestrzegających aktu, a w najpoważniejszych przypadkach może nałożyć na nich kary finansowe w wysokości do 6% globalnego obrotu⁴⁶⁸. Mechanizm egzekucyjny Komisji nie ogranicza się jednak do kar finansowych – koordynator ds. usług cyfrowych i Komisja mają prawo wymagać od usługodawców podjęcia natychmiastowych kroków, gdy zajdzie konieczność przeciwdziałania bardzo poważnym szkodom. W takich przypadkach platformy mogą przedstawić zobowiązania dotyczące sposobu ich naprawienia⁴⁶⁹.

Jeżeli Komisja podejrzewa, że dana VLOP naruszyła którykolwiek z przepisów DSA, może zdecydować o rozpoczęciu formalnego postępowania⁴⁷⁰. Komisja wszczęła już dochodzenie w sprawie X w związku z możliwym naruszeniem DSA w obszarach związanych z zarządzaniem ryzykiem, moderowaniem treści, przejrzystością reklam i dostępem do danych dla badaczy⁴⁷¹. W momencie sporządzania niniejszego raportu wspomniane dochodzenie jest w toku.

Jeżeli w trakcie postępowania Komisja stwierdzi naruszenie przepisów DSA, może podjąć dalsze kroki egzekucyjne, takie jak⁴⁷²:

- ❑ **Środki tymczasowe:** w przypadku pilnej potrzeby wynikającej z ryzyka wystąpienia poważnej szkody dla użytkowników_użytkowniczek, Komisja może zażądać podjęcia natychmiastowych działań w celu wyeliminowania takowej szkody. Aby ograniczyć takie ryzyko, wszelkie środki powinny być proporcjonalne i mieć charakter tymczasowy. Przykładowe środki tymczasowe mogą obejmować zmiany w systemach rekomendacji, wzmożone monitorowanie określonych słów kluczowych lub hashtagów oraz nakazy zaniechania lub usunięcia domniemanych naruszeń.
- ❑ **Zobowiązania wiążące:** platformy objęte postępowaniem egzekucyjnym mogą zobowiązać się wobec Komisji, że będą przestrzegać przepisów DSA. Jeżeli Komisja uzna te zobowiązania za skuteczne, może je zatwierdzić w formie oficjalnej decyzji.
- ❑ **Decyzja o nieprzestrzeganiu przepisów:** jeżeli Komisja stwierdzi, że naruszono przepisy DSA, zalecone środki tymczasowe lub podjęte zobowiązania, może wydać decyzję o nieprzestrzeganiu przepisów, po czym ma prawo nałożyć na przedsiębiorstwo grzywnę w wysokości do 1% jego rocznego obrotu.

⁴⁶⁸ Komisja Europejska, *Questions and answers on the Digital Services Act*.

⁴⁶⁹ Tamże.

⁴⁷⁰ Tamże.

⁴⁷¹ Komisja Europejska, *Commission opens formal proceedings against X under the Digital Services Act*, 18 grudnia 2023, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709.

⁴⁷² Komisja Europejska, *Questions and answers on the Digital Services Act*.

9. WNIOSKI I ZALECENIA

9.1 WNIOSKI

Niniejszy raport, oparty na wnikliwym dochodzeniu dotyczącym roli firmy X w tworzeniu warunków sprzyjających aktom TFGBV wobec społeczności LGBTI w Polsce, jednoznacznie wykazał, że firma przyczyniła się do naruszeń praw człowieka. X ponosi zatem odpowiedzialność za zapewnienie społeczności LGBTI w Polsce zadośćuczynienia oraz wdrożenie dodatkowych środków naprawczych w celu zapobiegnięcia powtórzeniu się takich naruszeń w przyszłości.

Jako że X jest przedsiębiorstwem globalnym, prowadzącym działalność w wielu krajach w UE i poza nią, istnieje znaczące ryzyko, że jego działania mogą sprzyjać dalszym aktom TFGBV, zwłaszcza w krajach nieanglojęzycznych. Dzieje się tak szczególnie ze względu na sprzężenie zbyt pobłażliwego podejścia platformy do moderowania treści oraz stosowania systemów rekomendacji opartych na zaangażowaniu użytkowników.

W kontekście nastrojów anti-LGBTI utrzymujących się w Polsce przez kilka ostatnich lat oraz faktu, że społeczność LGBTI znalazła się na celowniku rządu PiS, brak odpowiednich zasobów do moderowania treści oraz niedostateczna dbałość o przestrzeganie praw człowieka ze strony X przyczyniły się do normalizacji nienawiści, przemocy i dyskryminacji wobec tej społeczności. Pomimo deklaracji X o promowaniu idei „agory” dla wolności słowa, firma dopuściła do ataków o charakterze TFGBV na polską społeczność LGBTI, pozostawiając osobom LGBTI bardzo niewiele możliwości dochodzenia swoich praw. Jednocześnie firma konsekwentnie unika bezpośredniego kontaktu ze społecznością LGBTI, organizacjami społeczeństwa obywatelskiego zajmującymi się kwestiami LGBTI w Polsce, a nawet z polskim Stowarzyszeniem „Nigdy Więcej”, które jest zaufanym zgłaszającym X.

Wymienione uchybienia podsycaly nastroje anti-LGBTI i pozwoliły na coraz większą normalizację ataków na tę społeczność, co spowodowało, że X przyczynił się do TFGBV i naruszeń praw człowieka w nią wymierzonych.

Również w 2018 roku Amnesty International stwierdziła, że X (wówczas Twitter) nie wypełnia swoich obowiązków w zakresie ochrony praw człowieka dotyczących zapobiegania przemocy i nadużyciom wobec kobiet, w tym kobiet LGBTI, w swoim serwisie⁴⁷³. W 2020 roku Amnesty International ustaliła, że firma nie podjęła wystarczających kroków w celu przeciwdziałania występowaniu TFGBV na swojej platformie⁴⁷⁴. Powszechność obecności TFGBV na X świadczy o istnieniu systemowego problemu, który firma musi pilnie rozwiązać.

Fakt, że przez pięć lat od ostatniego dochodzenia Amnesty International X nadal przyczyniał się do TFGBV – oraz że firma radykalnie zmieniła swoje zasady dotyczące treści, przyjmując jeszcze bardziej pobłażliwe podejście do treści szkodliwych – wyraźnie świadczy o tym, że X zaniedbuje swoje obowiązki w zakresie ochrony praw człowieka. Budzi to wątpliwości co do tego, czy przed ogłoszeniem zmian w zasadach przeprowadzono jakąkolwiek adekwatną analizę należytej staranności w zakresie ochrony praw człowieka. Obawy dotyczą także gotowości firmy do podjęcia odpowiednich i skutecznych środków łagodzących, w tym zapewnienia wystarczających zasobów na moderowanie treści w językach innych niż angielski. Szczególnie niepokojące jest to, że X najwyraźniej nie uwzględnił w swojej najnowszej ocenie ryzyka na mocy DSA systemowych zagrożeń, jakie jego działalność stwarza dla społeczności LGBTI, a także nie potraktował poważnie zagrożeń dla praw innych niż prawo do wolności wypowiedzi.

Chociaż firma X wdrożyła kilka środków łagodzących, takich jak umożliwienie użytkownikom_ użytkownikom tworzenia list blokowanych osób oraz ograniczenie kręgu użytkowników_ użytkowniczek, którzy mogą odpowiadać na posty, zmiany są niewystarczające, aby odpowiednio złagodzić negatywny wpływ działalności firmy na prawa człowieka w Polsce. Ponadto podjęte środki mają zbyt ograniczony zakres, by

⁴⁷³ Amnesty International, *Toxic Twitter – a toxic place for women*.

⁴⁷⁴ Amnesty International, *Twitter's Scorecard*.

mogły zagwarantować niepowtórzenie się sytuacji, zwłaszcza że moderacja treści w języku polskim nadal jest niedostateczna.

Ryzyko, że X może przyczynić się do dalszych szkód wobec społeczności LGBTI w Polsce, wzrasta ze względu na model biznesowy platformy oparty na zaangażowaniu użytkowników_użytkowniczek. Z wpisów na blogu X i jego zasad polityki wynika, że zaangażowanie jest głównym filarem działalności platformy. W połączeniu z powszechnym występowaniem na niej treści anty-LGBTI prowadzi to do ryzyka algorytmicznego wzmacniania przypadków TFGBV, co należy uwzględnić przy wdrażaniu przyszłych środków ograniczających.

Brak właściwej oceny i ograniczenia zagrożenia dla społeczności LGBTI sugeruje, że X nie wypełnia swoich obowiązków nałożonych na przedsiębiorstwa przez postanowienia DSA, w szczególności tych opisanych w artykułach 34. i 35. Znajduje to potwierdzenie w niezależnym audycie oceny ryzyka przeprowadzonym przez X, który wykazał, że proces oceny ryzyka w firmie nie był wystarczająco rygorystyczny i że chociaż X wykazał świadomość działań związanych z zarządzaniem ryzykiem, zabrakło wystarczających dowodów na zaangażowanie firmy w podejmowanie decyzji w tym zakresie⁴⁷⁵. Niezwykle ważne jest, aby kolejna ocena ryzyka systemowego przeprowadzona przez X obejmowała dogłębną analizę i strategię ograniczania zagrożeń dla praw człowieka wykraczających poza wolność wypowiedzi oraz aby uwzględniała wynikające z jej działalności ryzyko naruszeń praw społeczności marginalizowanych, w szczególności osób LGBTI w Polsce, które doznały już krzywd w wyniku aktów TFGBV występujących na platformie.

Nieskuteczność X w ograniczaniu rozpowszechniania TFGBV na tej platformie od 2020 roku sugeruje brak chęci lub niezdolność firmy do wprowadzenia niezbędnych ulepszeń w celu złagodzenia spowodowanych szkód. X powinien niezwłocznie zmienić swoją strategię i dążyć do poprawy funkcjonowania, aby wywiązać się z obowiązku poszanowania praw człowieka, między innymi przeznaczając odpowiednie zasoby na moderowanie treści oraz nawiązując bardziej owocne stosunki z organizacjami społeczeństwa obywatelskiego w Polsce i we wszystkich krajach, w których prowadzi działalność.

Nieregulowany rozwój tak zwanego sektora Big Tech ma poważne konsekwencje dla praw człowieka społeczności marginalizowanych na całym świecie. Unia Europejska podjęła jednak kroki w celu przywrócenia równowagi w tej kwestii poprzez przyjęcie DSA. Dziś bardziej niż kiedykolwiek wcześniej ważne jest, aby Komisja Europejska, Parlament Europejski i państwa członkowskie UE wywiązywały się ze swojego obowiązku ochrony praw człowieka – w tym prawa do życia wolnego od przemocy ze względu na płeć – oraz rygorystycznie egzekwowały przepisy prawa. Niewystarczające moderowanie treści w połączeniu z modelem biznesowym opartym na inwigilacji sprawia, że X jest tykającą bombą zegarową. Stanowcze egzekwowanie przepisów DSA może zapobiec jej wybuchowi.

9.2 ZALECENIA

9.2.1 ZALECENIA DLA FIRMY X

ZADOŚĆUCZYNIENIE I ZAPOBIEGANIE PRZYSZŁYM SZKODOM W POLSCE

- ☐ Zapewnienie społeczności LGBTI w Polsce zadośćuczynienia za akty TFGBV, z którymi spotkała się ona w serwisie X. Zadośćuczynienie powinno przybrać formę publicznych przeprosin, a także zmiany praktyk moderowania treści i modelu biznesowego opartego na inwigilacji, co zagwarantuje, że sytuacja ta nie powtórzy się w przyszłości.
- ☐ Zreformowanie programu Trusted Flagger w Polsce, co da organizacjom społeczeństwa obywatelskiego i obrońcom praw człowieka większe możliwości udziału w procesie ustalania zasad publikowania treści.
- ☐ Istotne zwiększenie zdolności X do moderowania treści w języku polskim, w tym poprzez zatrudnienie większej liczby moderatorów mówiących po polsku i zapewnienie, że warunki ich pracy będą zgodne z normami praw człowieka.

⁴⁷⁵ FTI Consulting, *X Independent Audit*.

NALEŻYTA STARANNOŚĆ W PRZESTRZEGANIU PRAW CZŁOWIEKA

- ❑ Przeprowadzenie w X kompleksowego audytu i reorganizacji procesu należytej staranności w zakresie praw człowieka, z uwzględnieniem ochrony praw człowieka we wszystkich działaniach firmy/serwisu.
- ❑ Dopilnowanie, aby polityka i procesy należytej staranności w zakresie ochrony praw człowieka uwzględniały systemowe i powszechne oddziaływanie modelu biznesowego X na prawa człowieka jako takie, a także zwiększenie transparentności w zakresie sposobu identyfikacji oraz eliminowania zagrożeń i skutków działalności.
- ❑ Przeprowadzenie oceny wpływu działalności X na prawa człowieka w Polsce i pełne upublicznienie wyników takiej analizy.
- ❑ Zagwarantowanie, że projektowanie i wdrażanie systemów algorytmicznych na rynkach nieanglojęzycznych będzie obejmowało ocenę wpływu na prawa człowieka z uwzględnieniem merytorycznych konsultacji społecznych i zaangażowania społeczeństwa obywatelskiego, obrońców praw człowieka oraz przedstawicieli społeczności marginalizowanych lub niedostatecznie reprezentowanych przed sfinalizowaniem i wdrożeniem produktu lub usługi.

MODEL BIZNESOWY OPARTY NA INWIGILACJI

- ❑ Zaprzestanie gromadzenia danych osobowych o charakterze wrażliwym, jako że proceder ten narusza prawo do prywatności i zagraża szeregowi innych praw człowieka.
- ❑ Zaprzestanie stosowania reklamy ukierunkowanej i przyjęcie mniej szkodliwych alternatywnych modeli biznesowych, takich jak reklama kontekstowa.
- ❑ Poprawienie transparentności w stosowaniu algorytmów modelowania i moderowania treści przy zapewnieniu, że ich mechanizmy będą publicznie dostępne w zrozumiałej formie i regularnie aktualizowane, równoległe do wprowadzania zmian w systemach algorytmicznych.
- ❑ Umożliwienie niezależnym badaczom dostępu do danych będących przedmiotem interesu publicznego, w tym danych dotyczących moderowania treści i systemów algorytmicznych, oraz zapewnienie możliwości ich analizy.

MODERACJA TREŚCI

- ❑ Przeznaczenie stosownych zasobów na moderowanie treści w różnych językach na całym świecie ze szczególnym uwzględnieniem niwelowania istniejących nierówności, które w nieproporcjonalnym stopniu dotyczą kraje nieanglojęzyczne.
- ❑ Zapewnienie moderatorom treści szkoleń, w szczególności uwzględniających zagadnienie potrzeb osób LGBTI będących użytkownikami_użytkowniczkami platform oraz ryzyka związanego z TFGBV.
- ❑ Zapewnienie, aby wytyczne dotyczące moderacji treści opierały się na międzynarodowych przepisach i normach dotyczących praw człowieka, w tym normach dotyczących przeciwdziałania przemocy ze względu na płeć, oraz aby były z nimi zgodne.
- ❑ Zagwarantowanie, aby treści stanowiące akty TFGBV były objęte ograniczeniami zgodnymi z międzynarodowymi normami i prawami dotyczącymi praw człowieka, które dopuszczają ograniczenia wolności wypowiedzi w celu ochrony praw innych osób, pod warunkiem że ograniczenia te są konieczne i proporcjonalne.
- ❑ Zapewnienie, aby mechanizmy zgłaszania treści były odpowiednie, wystarczająco jasne, sprawne, szybkie i dostępne dla wszystkich użytkowników, także w popularnych językach innych niż angielski.
- ❑ Zapewnienie, aby weryfikacja faktów była odpowiednio finansowana, moderowanie treści nie było całkowicie powierzone funkcji „Uwagi społeczności”, a funkcja ta miała charakter uzupełniający w stosunku do moderowania treści, które winno być prowadzone zgodnie z polityką opartą na międzynarodowych przepisach i normach praw człowieka, a nie zastępować je.

- ☐ Natychmiastowy audyt zmian w polityce dotyczącej treści po 2022 roku i taka zmiana regulaminu, która pozwoli dostosować go do międzynarodowych przepisów i norm dotyczących praw człowieka, biorąc pod uwagę to, jak polityka dotycząca treści może wpłynąć na społeczności marginalizowane lub grupy wrażliwe.

9.2.1 ZALECENIA DLA KOMISJI EUROPEJSKIEJ

- ☐ Zapewnienie ścisłego egzekwowania przepisów Aktu o usługach cyfrowych, także poprzez dopilnowanie, aby państwa, które nie wyznaczyły koordynatorów ds. usług cyfrowych, uczyniły to w najbliższym możliwym terminie.
- ☐ Wprowadzenie wytycznych dla bardzo dużych platform internetowych do przeprowadzania oceny ryzyka systemowego, aby mieć pewność, że informacje zawarte w tych ocenach będą adekwatne i odpowiednie do wykorzystania przez społeczeństwo obywatelskie do kontroli tych platform.
- ☐ Rozszerzenie obecnych dochodzeń wszczętych przez Komisję Europejską w sprawie X o postępowanie dotyczące zdolności tego przedsiębiorstwa do skutecznego przeciwdziałania zagrożeniu TFGBV na platformie.
- ☐ Dopilnowanie, aby kary nakładane na przedsiębiorstwa w wyniku decyzji o nieprzestrzeganiu przepisów były wystarczająco wysokie, co zniechęci te przedsiębiorstwa do ponownego dopuszczenia się naruszeń.

9.2.2 ZALECENIA DLA RZĄDU RZECZPOSPOLITEJ POLSKIEJ

- ☐ Pełne wdrożenie Aktu o usługach cyfrowych; jak najszybsze powołanie koordynatora ds. usług cyfrowych, któremu będzie przysługiwać pełny zakres kompetencji.
- ☐ Dopilnowanie, aby po powołaniu koordynatora ds. usług cyfrowych urząd ten otrzymał odpowiednie zasoby merytoryczne, kadrowe i finansowe.

AMNESTY INTERNATIONAL TO GLOBALNY RUCH NA RZECZ PRAW CZŁOWIEKA. ZRZESZA LUDZI, KTÓRZY NIESPRAWIEDLIWOŚĆ NA ŚWIECIE TRAKTUJĄ OSOBIŚCIE.

W trakcie kampanii wyborczej w 2019 roku społeczność LGBTI stała się w Polsce głównym celem ataków ze strony polityków_polityczek i aktywistów_aktywistek sprzeciwiających się równouprawnieniu osób należących do tej społeczności. Wiele polityków_polityczek i aktywistów_aktywistek wykorzystywało X do szerzenia retoryki anty-LGBTI oraz podżegania do przemocy i dyskryminacji wobec tej społeczności.

Niniejszy raport zawiera analizę wpływu platformy X na szkody poniesione przez społeczność LGBTI w Polsce w latach 2019–2025. Ujawnia konsekwencje modelu biznesowego platformy X opartego na zaangażowaniu użytkowników_użytkowniczek, które negatywnie odbiły się na osobach LGBTI, w szczególności tych, które padły ofiarą mowy nienawiści na platformie. Raport wykazuje, że pomimo deklarowanego przez firmę dążenia do ochrony wolności słowa nadużycia online w serwisie przyczyniły się do tego, że społeczność LGBTI w Polsce żyje w strachu przed byciem sobą – zarówno w sieci, jak i poza nią.

Pomimo obowiązku ograniczenia ryzyka systemowego wynikającego z unijnego Aktu o usługach cyfrowych X nie podjął działań o należytej staranności w zakresie ochrony praw człowieka ani nie wdrożył środków ograniczających szkody wyrządzone w Polsce. Analiza Amnesty International jasno pokazuje, że model biznesowy tej platformy stanowi zagrożenie dla marginalizowanych społeczności na całym świecie. Podkreśla też konieczność rygorystycznego egzekwowania DSA i regulacji dotyczących technologii w skali globalnej, aby ochronić członków marginalizowanych społeczności przed negatywnymi skutkami takiego modelu biznesowego – niezależnie od tego, kim są i gdzie mieszkają.