RAPORT

Reakcja platform społecznościowych na zgłoszenia w I półroczu 2026 roku

ZESPÓŁ SZYBKIEGO REAGOWANIA I WYKRYWANIA DEZINFORMACJI



Ośrodek Analizy Dezinformacji
NASK

Reakcja platform społecznościowych na zgłoszenia w I półroczu 2026 roku

ZESPÓŁ SZYBKIEGO REAGOWANIA I WYKRYWANIA DEZINFORMACJI

Agnieszka Kaczmarczyk

Aleksandra Matusyk

Bartosz Popadiak

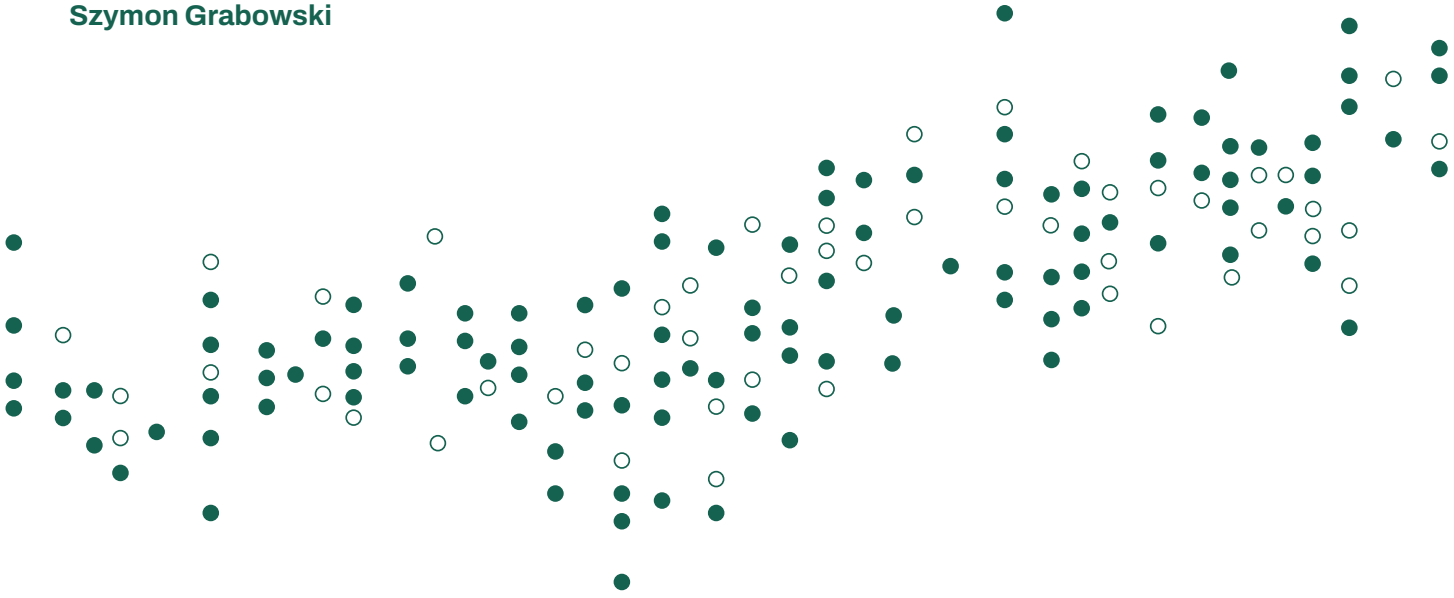
Dorota Kwiatkowska-Mroczek

Martyna Rataj

Mikołaj Bogusz

Piotr Muszyński

Szymon Grabowski



Ośrodek Analizy Dezinformacji NASK

NASK

PROJEKT FINANSOWANY ZE ŚRODKÓW MINISTERSTWA CYFRYZACJI

Tytuł
Reakcja platform społecznościowych na zgłoszenia w I półroczu 2026 roku.

Autorzy
Agnieszka Kaczmarczyk, Aleksandra Matusyk, Bartosz Popadiak,
Dorota Kwiatkowska-Mroczek, Martyna Rataj, Mikołaj Bogusz,
Piotr Muszyński, Szymon Grabowski

Redakcja i opieka merytoryczna
Ewelina Włodarczyk, Izabela Jarka

Korekta
Łukasz Szczęsny

Opracowanie graficzne
Marcin Ślusarczyk

Publikacja jest rozpowszechniana na zasadach licencji Creative Commons.
Uznanie autorstwa – Użycie niekomercyjne (CC BY-NC) 4.0 Międzynarodowe.

ISBN: 978-83-68356-78-6

2026

NASK – Państwowy Instytut Badawczy
ul. Kolska 12, 01-045 Warszawa
www.nask.pl

dezinformacja@nask.pl

Publikacja wyraża jedynie poglądy autorów i nie może być utożsamiana
z oficjalnym stanowiskiem Ministra Cyfryzacji

Spis treści

Wstęp 7

Metodologia obsługi zgłoszeń OAD NASK 9

- X.com (dawniej Twitter) 12
- Platformy Meta (Facebook, Instagram, Threads) 15
- TikTok 19
- Platformy Google (YouTube, Google Ads, Google Play) 23
- Telegram 26
- Discord 30
- LinkedIn 31
- Bluesky 31

Kluczowe wnioski 34

Podsumowanie pierwszego półrocza 2026 roku w podziale na platformy społecznościowe 38

Reakcja platform na wszystkie zgłoszenia OAD NASK 42

- X.com 45
- Platformy Meta 46
  - Facebook 46
  - Instagram 47
  - Threads 47
- TikTok 48
- Platformy Google (YouTube, Google Ads) 48

Telegram 49
Discord 49
Bluesky 50
Inne 50

Zgłoszenia wewnętrzne 52
X.com 53
Platformy Meta 54
Facebook 54
Instagram 55
Threads 55
TikTok 56
Platformy Google (YouTube, Google Ads) 56
Bluesky 57
Inne 57

Zgłoszenia zewnętrzne 59
X.com 60
Platformy Meta 61
Facebook 61
Instagram 62
Threads 62
TikTok 62
Platformy Google (YouTube, Google Ads) 63
Telegram 63
Discord 64
Inne 64

Case studies – narracje dezinformacyjne 66
Manipulowanie opinią publiczną i wpływ na postawy społeczne 67
Dezinformacja na temat dostaw agregatów do Ukrainy 67
Proirańska kampania dezinformacyjna z elementami CIB 71
Narracje dezinformacyjne wokół pożaru Puszczy Solskiej 76
Filmy AI na temat wypłat gotówki 79
Profile AI z polskimi żołnierzami 80

Monetyzacja dezinformacji 83
Konta Ukrainek wygenerowanych przy użyciu sztucznej inteligencji – fałszywe oferty matrymonialne 83
Fałszywi eksperci sprzedający suplementy 84
Kampanie reklamowe na platformach X i TikTok 86
Budowanie fałszywej wiarygodności 88
Profil „Straż Miejska Miasta Krakow” 88
Grupa na Facebooku „Iga Swiatek” 90

Podsumowanie: współpraca z platformami – trudności i wyzwania 95
X.com 95
Platformy Meta (Facebook, Instagram, Threads) 96
TikTok 102
Platformy Google (YouTube, Google Ads, Google Play) 102
Discord 104

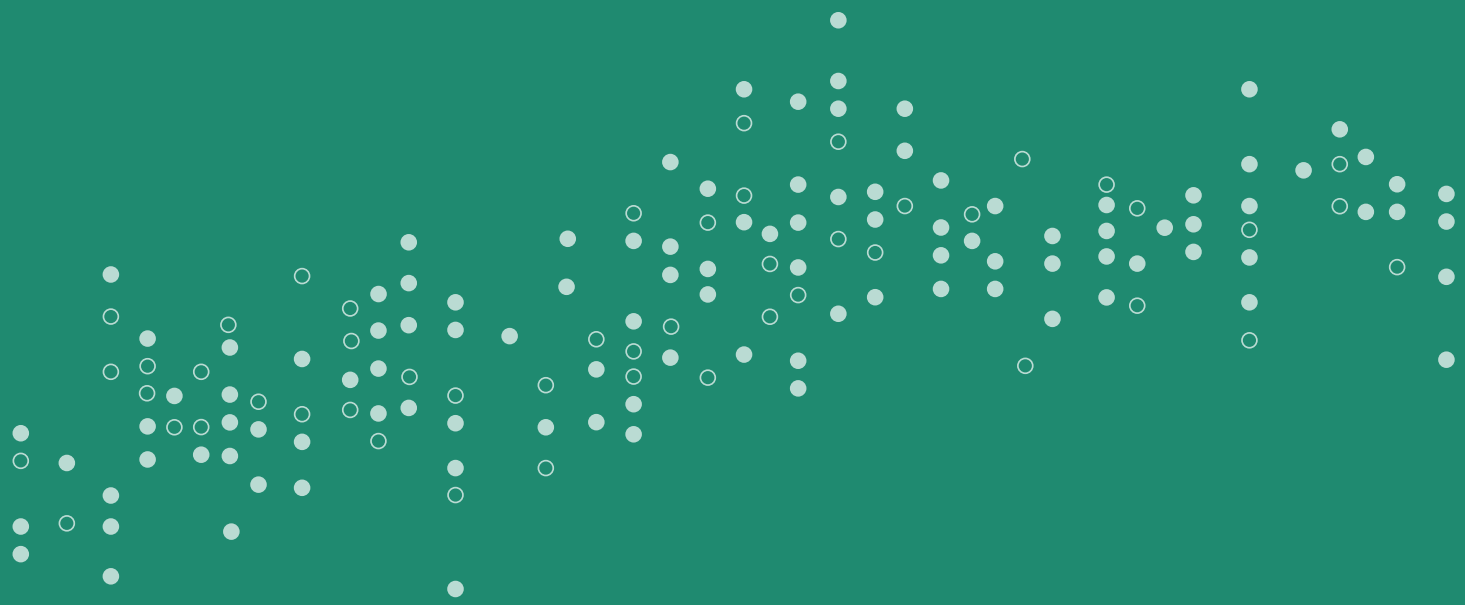
Wstęp

Raport przedstawia analizę działań podejmowanych przez platformy społecznościowe w odpowiedzi na zgłoszenia Ośrodka Analizy Dezinformacji (OAD NASK) w okresie od 1 stycznia do 30 czerwca 2026 roku.

W pierwszym półroczu 2026 roku zgłoszenia dotyczyły między innymi dezinformacji na temat dostaw agregatów do Ukrainy, proirańskich kampanii dezinformacyjnych zawierających elementy skoordynowanych działań nieautentycznych (*Coordinated Inauthentic Behaviour* – CIB), kont podszywających się pod ekspertów medycznych w celu sprzedaży suplementów, podszywania pod profile Straży Miejskiej, fałszywych narracji na temat pożaru w Puszczy Solskiej, a także kanałów, które prezentowały wygenerowane za pomocą sztucznej inteligencji rzekome wypowiedzi polskich żołnierzy. Ponadto odnotowano kampanie reklamowe wykorzystujące wizerunek powszechnie znanych osób. Wybrane przykłady opisano w rozdziale *Case studies – narracje dezinformacyjne*, dzieląc je ze względu na prawdopodobny cel wykorzystania poszczególnych narracji dezinformacyjnych.

Podsumowanie działań podejmowanych przez platformy w pierwszym półroczu 2026 roku wskazuje na wciąż niewystarczającą skuteczność mechanizmów moderacyjnych oraz zbyt niską aktywność administracji serwisów w zakresie przeciwdziałania dezinformacji. Analiza zgłoszeń przekazywanych platformom ujawnia powtarzalność problemów związanych z weryfikacją fałszywych i zmanipulowanych treści oraz opóźnionym reagowaniem na tego typu materiały. Niska efektywność procesów moderacyjnych sprzyja dalszemu rozpowszechnianiu dezinformacji i wywiera negatywny wpływ na debatę publiczną. Istotnym i stale aktualnym wyzwaniem pozostają również treści reklamowe wykorzystujące wizerunek instytucji publicznych, mediów oraz osób powszechnie znanych w celu dokonywania oszustw. Wyniki analiz z pierwszego półrocza wskazują na potrzebę dalszego wzmacniania mechanizmów weryfikacyjnych oraz usprawnienia procesów szybkiego identyfikowania i usuwania szkodliwych treści przez platformy internetowe.

Metodologia obsługi zgłoszeń OAD NASK



Metodologia obsługi zgłoszeń OAD NASK

Celem raportu jest przedstawienie reakcji platform społecznościowych na zgłoszone incydenty, a także identyfikacja głównych problemów i wyzwań w tym obszarze. W raporcie uwzględniono zgłoszenia wewnętrzne, pochodzące z codziennego monitoringu mediów społecznościowych prowadzonego przez OAD NASK przez siedem dni w tygodniu w systemie 24-godzinnym, a także zgłoszenia zewnętrzne, przesłane przez obywateli, instytucje bądź służby za pomocą zewnętrznego formularza na stronie zglos-dezinformacje.nask.pl oraz na skrzynkę e-mail dezinformacja@nask.pl.

Analiza obejmuje dane ilościowe dotyczące liczby wszystkich zgłoszeń, treści usuniętych, materiałów poddanych moderacji oraz zgłoszeń odrzuconych przez platformy społecznościowe. W raporcie uwzględniono także szczegółową kategoryzację incydentów, która określa rodzaj naruszenia zidentyfikowany na podstawie regulaminów platform społecznościowych bądź innych regulacji prawnych. Zgodnie z nimi przyjęto wewnętrzny podział na trzy główne kategorie: *fałszywa informacja; manipulacja; podszywanie pod instytucję lub osobę powszechnie znaną*. Kategoryzacja nie jest bezpośrednim odzwierciedleniem regulaminów, jednak pozwala na zachowanie spójności między kwalifikacją incydentów przez OAD NASK, a kategoryzacją wykorzystywaną przez serwisy społecznościowe w swoich formularzach i narzędziach zgłoszeniowych.

Raport zawiera również *case studies*, które OAD NASK sklasyfikował i podzielił na trzy typy ze względu na prawdopodobny cel realizowanej dezinformacji: *manipulowanie opinią publiczną i wpływ na postawy społeczne; monetyzacja dezinformacji; budowanie fałszywej wiarygodności.*

W pierwszym półroczu opisane szczegółowo przypadki dotyczyły między innymi następujących tematów:

- dezinformacji na temat dostaw agregatów do Ukrainy,
- proirańskiej kampanii dezinformacyjnej z elementami CIB,
- kont fałszywych ekspertów sprzedających suplementy,
- kont z materiałami wygenerowanymi przy użyciu AI przedstawiających fałszywy wizerunek Ukrainek,
- dezinformacji na temat pożaru w Puszczy Solskiej,
- materiałów AI przedstawiających rzekome wypowiedzi polskich żołnierzy,
- oszustw w postaci reklam, w których podszywano się pod znane media,
- publicznej grupy, na której publikowano treści dezinformujące,
- podszywania pod profile polskich służb.

Ilustrują one specyfikę zjawiska dezinformacji i prezentują reakcje platform na zgłoszenia w raportowanym okresie. Przypadki, które zostały omówione szczegółowo, dobrano na podstawie bieżących obserwacji analityków i analityczek OAD NASK.

Na potrzeby raportu należy wyjaśnić przyjętą definicję zgłoszenia. Pojęcie to jest stosowane w opracowaniu wymiennie z incydentem dezinformacyjnym. Określa wszystkie materiały zidentyfikowane przez OAD NASK, które noszą znamiona naruszenia zapisów regulaminów platform społecznościowych lub łamią inne regulacje, na przykład prawo krajowe bądź unijne. Tylko zgłoszenia tego rodzaju materiałów uznawane są za zasadne i przekazywane przez OAD NASK do punktów kontaktowych platform społecznościowych lub za pomocą dedykowanych narzędzi zgłoszeniowych oferowanych OAD NASK przez administracje poszczególnych serwisów. Ze względu na swoje pochodzenie zgłoszenia dzielą się na wewnętrzne i zewnętrzne, zgodnie z wcześniej wyjaśnioną ścieżką.

Głównym obszarem obsługi są incydenty dezinformacyjne, których treść jest stworzona w języku polskim. Wyjątek stanowią zgłoszenia dotyczące szerszych kampanii dezinformacyjnych, takich jak Undercut czy Overload.

W zależności od reakcji platform społecznościowych zgłoszenia dzielą się na:

A. **odrzucone**, czyli uznane przez zespół moderacji za niestanowiące naruszenia regulaminu bądź takie, które w ciągu dwóch tygodni nie uzyskały żadnej odpowiedzi. W przypadku zgłoszeń odrzuconych OAD NASK zawsze dokonuje odwołania przez dedykowany punkt kontaktowy, za pomocą udostępnionego narzędzia lub, jeśli to możliwe, z poziomu użytkownika;

B. **poddane moderacji**, czyli oznaczone w jakikolwiek widoczny dla użytkowników i użytkowniczek sposób jako treści dezinformacyjne/manipulacyjne, na przykład za pomocą systemu community notes, przez umieszczenie stosownej nakładki informacyjnej na materiale lub częściowe usunięcie fragmentu wideo, również wpisy/konta zawieszone;

C. **usunięte**, czyli takie, które zostały w pełni obsłużone, a wskazana treść lub całe konto/kanał zostały zablokowane albo trwale usunięte.

Co istotne, platformy społecznościowe nie informują o przewidywanym czasie rozpatrzenia zgłoszenia. Obejmuje to również te serwisy, które udostępniają Ośrodkowi specjalne narzędzia przeznaczone do obsługi zgłoszeń. Z tego względu OAD NASK nie jest w stanie w krótkim czasie udzielić odpowiedzi obywatelom bądź instytucjom, od których pochodzą zgłoszenia, ponieważ nie zna realnego terminu, w którym należy spodziewać się odpowiedzi platform. Każde zgłoszenie musi być śledzone indywidualnie przez osobę z zespołu analitycznego. W przypadku braku reakcji ze strony platformy zgłoszenie musi być ponawiane.

W raporcie omówiono także wyzwania wynikające bezpośrednio z charakterystyki współpracy z platformami społecznościowymi w zakresie obsługi zgłoszeń, takie jak brak przejrzystych procedur, długi czas reakcji czy niekonsekwentne podejście do moderacji treści.

Poniżej przedstawiono szczegółowo proces obsługi przez OAD NASK zgłoszeń do administracji poszczególnych platform społecznościowych wraz z rodzajami naruszeń odpowiadających regulaminom tych serwisów.

X.com (dawniej Twitter)

Proces zgłaszania treści dezinformacyjnych

Analitycy i analityczki OAD NASK mogą zgłaszać treści dezinformacyjne za pomocą dwóch udostępnionych przez platformę X ścieżek.

A. Formularz zgłoszeniowy, który jest dołączony do każdego materiału opublikowanego w serwisie i pozostaje dostępny dla wszystkich użytkowników i użytkowniczek platformy. Osoby zgłaszające naruszenie mają możliwość przypisania incydentu do jednej z określonych kategorii:

- **a. spam**
  fałszywe interakcje, oszustwa; fałszywe konta, złośliwe linki,
- **b. nienawiść**
  obelgi, rasistowskie lub seksistowskie stereotypy, dehumanizacja, wzbudzanie strachu lub podżeganie do dyskryminacji, nawiązania, symbole i logo nacechowane nienawiścią,
- **c. nadużycia i nękanie**
  obelgi, niechciane treści o zabarwieniu erotycznym i drastyczne treści sugerujące uprzedmiotowienie innych osób, niechciane treści nie do oglądania w pracy lub drastyczne; negowanie aktów przemocy, nękanie konkretnych osób i podżeganie do nękania,
- **d. bezpieczeństwo dzieci**
  wykorzystywanie seksualne dzieci, grooming, przemoc fizyczna wobec dzieci, zbyt młody użytkownik,
- **e. mowa nienawiści**
  groźby karalne, życzenia obrażeń ciała, gloryfikacja przemocy, podżeganie do przemocy, zawoalowane podżeganie do przemocy,
- **f. multimedia drastyczne lub przedstawiające przemoc**
  drastyczne treści, nieuzasadnione okrucieństwo, nagość i zachowania seksualne osób dorosłych, brutalne zachowania na tle seksualnym, zoofilia i nekrofilia, multimedia przedstawiające zmarłą osobę,
- **g. zachowania nielegalne i regulowane przepisami prawa**
  handel ludźmi, usługi seksualne, narkotyki, broń, zagrożone gatunki, podejmowanie działań niezgodnych z prawem,
- **h. naśladownictwo**
  udawanie, że jest się kimś innym; obejmuje to także (w teorii) niezgodne konta parodystyczne/fanowskie,

i. **treści pornograficzne**
publikowanie, rozpowszechnianie lub promowanie materiałów o charakterze pornograficznym bez zgody osób zaangażowanych, w tym treści intymnych lub seksualnych, takich jak nagrania z ukrytej kamery, zdjęcia typu upskirt, deepfake'i z nagim wizerunkiem, prywatne materiały z intymnych sytuacji oraz oferowanie nagrody za pozyskanie takich materiałów.

j. **treści prywatne albo utworzone bez zgody lub wiedzy osób, których dotyczą**
udostępnianie informacji osobistych, grożenie udostępnieniem/ujawnieniem informacji osobistych, udostępnianie obrazów przedstawiających sytuacje intymne, na które nie wyrażono zgody; udostępnianie przedstawiających mój wizerunek obrazów, których nie życzę sobie na platformie,

k. **samobójstwo lub samookaleczenie**
zachęcanie do dokonywania aktów samookaleczenia, rozpowszechnianie takich zachowań lub udzielanie porad lub instrukcji ich dotyczących,

l. **terroryzm lub ekstremizm**
skrajny ekstremizm i terroryzm, grupy oraz sieci propagujące nienawiść,

m. **uczciwość obywatelska**
wprowadzające w błąd treści dotyczące uczestnictwa, zniechęcania do uczestnictwa lub zastraszania osób głosujących w wyborach albo uczestniczących w innych procedurach obywatelskich.

Udostępniona została również dodatkowa klasyfikacja na potrzeby zgłaszania nielegalnych treści opisanych w Akcie o usługach cyfrowych (*Digital Services Act*, DSA)[1]. Wśród dostępnych kategorii znajdują się:

a. dobrostan zwierząt,

b. naruszenie praw konsumenta dotyczących informacji,

c. przemoc w cyberprzestrzeni,

d. naruszenia zasad dotyczących ochrony danych i prywatności,

e. nielegalna lub szkodliwa mowa,

f. negatywne skutki dla dyskursu obywatelskiego lub wyborów,

1 https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (dostęp: 12.09.2026 godz. 09:50)

g. nieakceptowalne zachowanie,

h. pornografia lub treści o charakterze seksualnym,

i. ochrona nieletnich,

j. ryzyko dla bezpieczeństwa publicznego,

k. oszustwa i wyłudzenia,

l. zakres usług platformy,

m. samookaleczenie,

n. produkty niebezpieczne i nielegalne,

o. przemoc.

B. Analitycy i analityczki OAD NASK posiadają również dostęp do narzędzia zgłoszeniowego *Legal request submissions*. Narzędzie to dedykowane jest między innymi funkcjonariuszom i funkcjonariuszkom organów ścigania oraz urzędnikom i urzędniczkom państwowym. Umożliwia ono złożenie wniosku o wstrzymanie publikacji treści na podstawie odpowiedniego żądania prawnego, w tym także w przypadku stwierdzenia występowania treści o charakterze dezinformacyjnym.

Zgodnie z ustaleniami poczynionymi między administracją platformy X a zespołem OAD NASK zdecydowano, że wymienione poniżej typy incydentów nie wymagają zgłoszenia za pośrednictwem wymienionych narzędzi, a informacja o nich zostaje przekazana za pomocą wiadomości e-mail bezpośrednio do punktu kontaktowego wyznaczonego przez administrację:

- treści typu deepfake i inne rodzaje manipulacji materiałami audiowizualnymi,
- siatki kont-botów oraz pozostałe skoordynowane działania nieautentyczne (CIB).

Platformy Meta (Facebook, Instagram, Threads)

W odpowiedzi na rosnące zagrożenie związane z dezinformacją firma Meta udostępnia pracownikom i pracownicom OAD NASK dedykowaną platformę, która umożliwia zgłaszanie szkodliwych treści z serwisów Facebook, Instagram oraz Threads. Narzędzie pozwala na śledzenie bieżącego statusu zgłoszeń oraz przeglądanie historii wcześniejszych spraw, a także na złożenie odwołania w przypadku odrzucenia zgłoszenia.

Zespołowi zapewniono również wsparcie ze strony polskojęzycznego opiekuna, reagującego w przypadku pilnych lub szczególnie groźnych incydentów. Przedstawiciel ten reprezentuje jednak cały region Europy Wschodniej i Środkowej, a więc czas reakcji bywa wydłużony. Dodatkowym ograniczeniem we współpracy jest fakt, że opiekun nie posiada uprawnień do samodzielnego rozpatrywania zgłoszeń – jego rolą jest jedynie przekazywanie zgłoszeń do zespołu moderacji.

Na platformie Facebook dostępny jest formularz zgłoszeniowy, który jest dołączony do każdego materiału opublikowanego w serwisie i pozostaje dostępny dla wszystkich użytkowników i użytkowniczek platformy. Osoby zgłaszające naruszenie mają możliwość przypisania incydentu do jednej z określonych kategorii poniżej.

Kategorie zgłoszeń:

a. **problem związany z osobą poniżej 18 roku życia**
   treści dotyczące wykorzystywania, krzywdzenia, narażania lub nieodpowiedniego angażowania osób niepełnoletnich.

b. **przemoc, prześladowanie lub nękanie**
   groźby, zastraszanie, nękanie, prześladowanie, publikowanie treści mających na celu skrzywdzenie lub upokorzenie konkretnej osoby.

c. **samobójstwo lub samookaleczenie**
   promowanie, zachęcanie, gloryfikowanie lub przedstawianie samobójstwa i samookaleczeń.

d. **brutalne, nienawistne lub niepokojące treści**
   treści przedstawiające przemoc, okrucieństwo, mowę nienawiści, treści drastyczne lub wywołujące silny dyskomfort.

e. **sprzedaż lub promocja produktów podlegających ograniczeniom**
handel lub promocja produktów i usług objętych ograniczeniami, takich jak alkohol, wyroby tytoniowe, broń lub inne regulowane towary.

f. **zawartość dla dorosłych**
publikowanie, rozpowszechnianie lub promowanie materiałów o charakterze pornograficznym lub seksualnym, w tym treści intymnych udostępnianych bez zgody osób zaangażowanych, takich jak nagrania z ukrytej kamery, zdjęcia typu upskirt, cyfrowo zmanipulowane nagie wizerunki (deepfake), prywatne materiały z intymnych sytuacji oraz oferowanie nagrody za pozyskanie takich materiałów.

g. **oszustwo lub fałszywe informacje**
oszustwa, wyłudzanie danych lub pieniędzy, podszywanie się, wprowadzające w błąd informacje oraz rozpowszechnianie fałszywych treści.

h. **nie chcę tego widzieć**
treści niepożądane przez użytkownika, które nie muszą naruszać zasad platformy, ale które użytkownik chce ukryć lub ograniczyć.

i. **zawartość niezgodna z prawem**
treści naruszające obowiązujące przepisy prawa, w tym działalność przestępczą, handel nielegalnymi towarami lub inne działania zabronione prawnie.

Podobny formularz dostępny jest na platformie Instagram, jednak kategorie zgłoszeń różnią się pomiędzy platformami. Formularz również dołączony jest do każdego materiału opublikowanego w serwisie i pozostaje dostępny dla wszystkich użytkowników i użytkowniczek platformy. Osoby zgłaszające naruszenie mają możliwość przypisania incydentu do jednej z określonych kategorii poniżej:

a. **po prostu mi się nie podoba**
treści, które użytkownik uważa za nieinteresujące, nieodpowiednie lub których nie chce oglądać, mimo że mogą nie naruszać zasad platformy.

b. **nękanie lub niechciane kontakty**
prześladowanie, zastraszanie, groźby, uporczywe wiadomości, niechciane interakcje oraz działania wymierzone w konkretną osobę.

c. **samobójstwo, zachowania autodestrukcyjne lub zaburzenia odżywiania**
promowanie, zachęcanie, gloryfikowanie lub przedstawianie samobójstwa, samookaleczeń, zaburzeń odżywiania oraz innych zachowań autodestrukcyjnych.

d. **przemoc, nienawiść lub wykorzystywanie**
treści przedstawiające lub promujące przemoc, mowę nienawiści, dyskryminację, wykorzystywanie ludzi lub zwierząt oraz inne formy krzywdzenia.

e. **sprzedaż lub promocja produktów podlegających ograniczeniom**
handel, reklama lub promocja produktów i usług objętych ograniczeniami, takich jak alkohol, wyroby tytoniowe, broń, substancje psychoaktywne lub inne regulowane towary.

f. **nagość lub aktywność seksualna**
publikowanie, rozpowszechnianie lub promowanie treści zawierających nagość, czynności seksualne, materiały pornograficzne lub inne treści o charakterze seksualnym, w tym treści intymnych udostępnianych bez zgody osób zaangażowanych.

g. **scam, oszustwo lub spam**
oszustwa, wyłudzanie danych lub pieniędzy, podszywanie się pod inne osoby, masowe niechciane wiadomości oraz inne zwodnicze praktyki.

h. **fałszywe informacje**
publikowanie lub rozpowszechnianie treści, które mogą wprowadzać odbiorców w błąd poprzez podawanie nieprawdziwych, zmanipulowanych lub niezweryfikowanych informacji.

i. **zawartość niezgodna z prawem**
treści naruszające obowiązujące przepisy prawa, w tym działalność przestępczą, handel nielegalnymi towarami, naruszenia praw własności intelektualnej lub inne działania zabronione prawnie.

Formularz zgłoszeniowy do platformy Threads znacząco różni się od innych formularzy w ramach grupy Meta, a jego kategorie obejmują:

a. **własność intelektualna**
naruszenia praw własności intelektualnej, w tym praw autorskich, znaków towarowych, patentów lub innych chronionych praw przysługujących osobom fizycznym lub podmiotom.

b. **zniesławienie**
publikowanie treści zawierających nieprawdziwe informacje, które mogą naruszać dobre imię, reputację lub wiarygodność osoby fizycznej, organizacji albo podmiotu.

c. **prywatność / usunięcie zgodnie z RODO**
ujawnianie danych osobowych lub innych informacji naruszających prywatność, a także treści podlegające usunięciu na podstawie przepisów dotyczących ochrony danych osobowych, w tym RODO.

d. **naruszenie innych przepisów prawa**
treści, które mogą naruszać przepisy niewymienione w pozostałych kategoriach, np. mowę nienawiści, nawoływanie do popełnienia przestępstwa lub inne działania niezgodne z obowiązującym prawem.

e. **elections** (przyp. red. wybory)
treści związane z procesami wyborczymi, które mogą naruszać lokalne przepisy dotyczące wyborów, kampanii politycznych, finansowania kampanii lub rozpowszechniania informacji wyborczych.

TikTok

W przypadku identyfikacji treści o charakterze dezinformacyjnym w serwisie TikTok analitycy i analityczki OAD NASK korzystają z dedykowanego narzędzia zgłoszeniowego. Umożliwia ono zarówno dodawanie nowych zgłoszeń, jak i przeglądanie już zarejestrowanych spraw, przy czym nie przewidziano w nim funkcjonalności umożliwiającej odwołanie się od decyzji odmownej. Narzędzie to funkcjonuje w języku angielskim. W przypadku zgłoszeń o szczególnym znaczeniu zespół ma możliwość skorzystania ze wsparcia polskojęzycznego opiekuna pełniącego rolę punktu kontaktowego.

Na platformie TikTok dostępne jest narzędzie do zgłaszania treści z poziomu użytkownika, obejmuje ono poniższe kategorie:

a. **Przemoc, nadużycia i wykorzystywanie w celach przestępczych**
treści promujące, przedstawiające lub organizujące przemoc, wykorzystywanie ludzi oraz działalność przestępczą.

b. **Nienawiść i prześladowanie**
mowa nienawiści, dyskryminacja, zastraszanie, nękanie oraz inne formy atakowania osób lub grup.

c. **Samobójstwo i samookaleczanie**
promowanie, zachęcanie, gloryfikowanie lub przedstawianie samobójstwa oraz samookaleczeń.

d. **Zaburzone odżywianie i niezdrowy obraz ciała**
treści promujące zaburzenia odżywiania, niebezpieczne praktyki żywieniowe lub nierealistyczne wzorce dotyczące wyglądu.

e. **Niebezpieczne działania i wyzwania**
zachęcanie do podejmowania ryzykownych, szkodliwych lub potencjalnie niebezpiecznych działań i wyzwań.

f. **Nagość i treści seksualne**
publikowanie, rozpowszechnianie lub promowanie treści zawierających nagość, aktywność seksualną, pornografię lub inne materiały o charakterze seksualnym.

g. **Treści szokujące i drastyczne**
materiały przedstawiające brutalną przemoc, okrucieństwo, poważne obrażenia, śmierć lub inne drastyczne sceny.

h. **Błędne informacje**
publikowanie lub rozpowszechnianie nieprawdziwych, zmanipulowanych lub wprowadzających w błąd informacji.

i. **Zwodnicze zachowania i spam**
spam, manipulowanie użytkownikami, podszywanie się pod inne osoby oraz inne działania mające na celu wprowadzenie odbiorców w błąd.

j. **Towary i działalność regulowana**
handel, promocja lub dystrybucja produktów i usług podlegających ograniczeniom prawnym, takich jak alkohol, wyroby tytoniowe, broń, narkotyki czy usługi finansowe.

k. **Oszustwa i nadużycia finansowe**
wyłudzanie pieniędzy lub danych, phishing, scamy inwestycyjne oraz inne formy oszustw finansowych.

l. **Udostępnianie danych osobowych**
publikowanie lub rozpowszechnianie prywatnych danych osobowych bez zgody zainteresowanej osoby.

m. **Treści niezgodne z prawem**
materiały naruszające obowiązujące przepisy prawa, w tym działania przestępcze lub inne nielegalne aktywności.

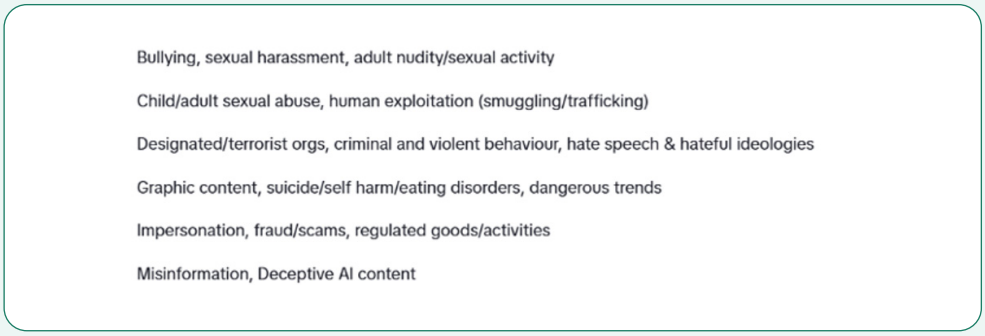
n. **Podróbki i własność intelektualna**
sprzedaż lub promowanie podróbek oraz naruszanie praw autorskich, znaków towarowych i innych praw własności intelektualnej.

o. **Ukryte treści markowe**
nieujawniona reklama, lokowanie produktów lub treści sponsorowane, które nie zostały odpowiednio oznaczone.

p. **Inne**
zgłoszenia dotyczące treści, które nie mieszczą się w żadnej z powyższych kategorii.

W omawianym okresie na platformie zgłoszeniowej TikTok pojawiły się nowe kategorie naruszeń. Wcześniej przy zgłaszaniu treści dostępna była jedynie ogólna kategoria do oznaczenia materiałów „łamiących zasady platformy", bez podziału na poszczególne rodzaje naruszeń.

OBRAZ 1. Nowe kategorie naruszeń dostępne na platformie zgłoszeniowej TikTok



ŹRÓDŁO: TIKTOK.COM

Kategorie zgłoszeń:

a. **Bullying, Sexual Harassment, Adult Nudity/Sexual Activity (Nękanie, molestowanie seksualne, nagość lub aktywność seksualna osób dorosłych)**
obraźliwe lub nękające zachowania, molestowanie seksualne, a także treści zawierające nagość, aktywność seksualną lub materiały o charakterze erotycznym z udziałem osób dorosłych.

b. **Child/Adult Sexual Abuse, Human Exploitation (Smuggling/ Trafficking) (Wykorzystywanie seksualne dzieci lub dorosłych, wyzysk ludzi, przemyt lub handel ludźmi)**
treści związane z wykorzystywaniem seksualnym, przemocą seksualną, handlem ludźmi, przemytem osób, pracą przymusową lub innymi formami wyzysku.

c. **Designated/Terrorist Orgs, Criminal and Violent Behaviour, Hate Speech & Hateful Ideologies (Organizacje terrorystyczne, działalność przestępcza i brutalne zachowania, mowa nienawiści oraz ideologie nienawistne)**
promowanie lub wspieranie organizacji terrorystycznych, działalności przestępczej, aktów przemocy, mowy nienawiści, dyskryminacji oraz ideologii opartych na nienawiści.

d. **Graphic Content, Suicide/Self Harm/Eating Disorders, Dangerous Trends (Treści drastyczne, samobójstwo, samookaleczenia, zaburzenia odżywiania oraz niebezpieczne trendy)**
materiały przedstawiające brutalną przemoc, okrucieństwo, samobójstwo, samookaleczenia, zaburzenia odżywiania lub zachęcające do udziału w niebezpiecznych działaniach i wyzwaniach.

e. **Impersonation, Fraud/Scams, Regulated Goods/Activities (Podszywanie się, oszustwa i scamy, produkty lub działania regulowane przepisami)**
podszywanie się pod inne osoby lub podmioty, oszustwa finansowe, phishing, wyłudzanie danych, scamy oraz handel lub promocja produktów i usług podlegających ograniczeniom prawnym.

f. **Misinformation, Deceptive AI Content (Dezinformacja i wprowadzające w błąd treści wygenerowane przez sztuczną inteligencję)**
publikowanie fałszywych lub zmanipulowanych informacji oraz treści wygenerowanych przez sztuczną inteligencję, które mogą wprowadzać odbiorców w błąd co do ich autentyczności lub przedstawianych faktów.

Platformy Google (YouTube, Google Ads, Google Play)

W przypadku platform i usług zarządzanych przez Google analitycy i analityczki OAD NASK mają możliwość zgłaszania treści za pośrednictwem ogólnodostępnych formularzy dedykowanych serwisom YouTube, Google Play oraz Google Ads. Zgłoszenia dotyczące szkodliwych materiałów występujących na pozostałych platformach zarządzanych przez Google są przekazywane drogą e-mailową do polskojęzycznego punktu kontaktowego.

Ogólnodostępny formularz zgłoszeniowy podzielony jest na kategorie:

a. **Treści o charakterze seksualnym**
publikowanie, rozpowszechnianie lub promowanie treści zawierających nagość, aktywność seksualną, pornografię lub inne materiały o charakterze seksualnym.

b. **Treści przedstawiające przemoc lub budzące odrazę**
materiały przedstawiające brutalną przemoc, okrucieństwo, poważne obrażenia, śmierć lub inne drastyczne sceny mogące wywoływać szok lub odrazę.

c. **Treści krzywdzące lub szerzące nienawiść**
mowa nienawiści, dyskryminacja, poniżanie, dehumanizacja oraz inne treści atakujące osoby lub grupy ze względu na ich cechy chronione.

d. **Nękanie lub dokuczanie**
zastraszanie, groźby, prześladowanie, obraźliwe komentarze oraz inne działania wymierzone w konkretną osobę.

e. **Sceny szkodliwe lub niebezpieczne**
treści promujące niebezpieczne zachowania, ryzykowne wyzwania lub działania mogące prowadzić do szkody fizycznej albo psychicznej.

f. **Samobójstwo, samookaleczanie lub zaburzenia odżywiania**
promowanie, zachęcanie lub gloryfikowanie samobójstwa, samookaleczeń oraz zaburzeń odżywiania.

g. **Nieprawdziwe informacje**
publikowanie lub rozpowszechnianie fałszywych, zmanipulowanych lub wprowadzających w błąd informacji.

h. **Znęcanie się nad dziećmi**
treści przedstawiające, promujące lub wspierające wykorzystywanie, krzywdzenie lub narażanie dzieci na niebezpieczeństwo.

i. **Propagowanie terroryzmu**
wspieranie, promowanie, rekrutowanie lub gloryfikowanie organizacji terrorystycznych, ideologii terrorystycznych lub aktów terroru.

j. **Spam lub treści wprowadzające w błąd**
masowe niechciane treści, oszustwa, manipulacje oraz działania mające na celu wprowadzenie użytkowników w błąd.

k. **Problem prawny**
treści mogące naruszać przepisy prawa, prawa autorskie, prywatność, dobra osobiste lub inne regulacje prawne.

W ramach usługi Google Ads dostępny jest formularz zgłoszeniowy zawierający poniżej wypisane kategorię naruszeń:

a. **Wprowadzanie w błąd lub oszustwo**
nieprawdziwe twierdzenia i oferty, podszywanie się, phishing, clickbait oraz błędne informacje o cenach.

b. **Szkodliwe, drastyczne lub niebezpieczne**
nienawiść, nękanie w internecie, samookaleczanie, dyskryminacja, okrucieństwo wobec zwierząt, używanie i promowanie narkotyków.

c. **Treści o charakterze seksualnym**
nagość, spotkania w celach seksualnych, gadżety seksualne, kluby ze striptizem, czaty na żywo z podtekstem seksualnym, erotyczne piosenki.

d. **Promowanie produktów lub usług objętych ograniczeniami**
alkohol, tytoń, hazard, usługi związane z leczeniem uzależnień, opieka zdrowotna, polityka, fałsze.

e. **Narusza konkretne przepisy lub moje prawa**
narusza przepisy obowiązujące w kraju bądź regionie lub jestem właścicielem praw autorskich bądź znaku towarowego i mam zastrzeżenia dotyczące własności intelektualnej.

f. **Coś innego**
zgłoszenia dotyczące treści, które nie mieszczą się w żadnej z powyższych kategorii.

W przypadku zgłoszeń w ramach usługi Google Play dostępny jest formularz zgłoszeniowy z kategoriami:

a. **Misinformation (Dezinformacja)**
publikowanie lub rozpowszechnianie fałszywych, niezweryfikowanych lub wprowadzających w błąd informacji, które mogą zniekształcać odbiór faktów.

b. **Manipulated Media (Zmanipulowane materiały medialne)**
zdjęcia, nagrania audio lub wideo, które zostały zmodyfikowane w sposób mogący wprowadzać odbiorców w błąd co do ich autentyczności lub przedstawianych wydarzeń.

c. **Deceptive Practices (Zwodnicze praktyki)**
działania mające na celu oszukiwanie użytkowników, ukrywanie rzeczywistych intencji, manipulowanie odbiorcami lub osiąganie korzyści poprzez wprowadzanie w błąd.

d. **Misleading Content (Treści wprowadzające w błąd)**
treści prezentowane w sposób mogący prowadzić do błędnych wniosków, nawet jeśli nie zawierają bezpośrednio fałszywych informacji.

e. **Harmful Conspiracy Theories (Szkodliwe teorie spiskowe)**
treści promujące lub rozpowszechniające teorie spiskowe, które mogą prowadzić do dezinformacji, szkód społecznych, narażania zdrowia lub bezpieczeństwa użytkowników.

Telegram

Administracja serwisu Telegram nie udostępnia zespołowi OAD dedykowanego narzędzia umożliwiającego zgłaszanie treści o charakterze dezinformacyjnym. Zgłoszenia mogą być przekazywane wyłącznie za pośrednictwem ogólnodostępnego formularza lub poprzez e-mail na adres abuse@telegram.com. Po jego wypełnieniu osoba zgłaszająca nadużycie nie otrzymuje żadnego potwierdzenia przyjęcia zgłoszenia, w tym numeru referencyjnego ani informacji o dalszym statusie sprawy.

W ramach formularza we wszystkich kategoriach oprócz „Nie podoba mi się" oraz „To nie jest nielegalne, ale musi zostać usunięte" użytkownik ma możliwość przedstawienia szczegółowego opisu zgłoszenia. We wspomnianych kategoriach natomiast brak jest opcji dodania dodatkowych informacji – wybór jednej z nich skutkuje automatycznym przekazaniem zgłoszenia do zespołu moderującego.

Platforma umożliwia również zgłaszanie nielegalnych treści za pośrednictwem czatbota, jednak jego funkcjonalność pozostaje ograniczona. W praktyce wybór tego narzędzia oznacza, że możliwe jest zgłoszenie wyłącznie pojedynczego materiału, bez opcji wskazania całego konta. Dotychczasowe doświadczenia analityków i analityczek wskazują ponadto na niską skuteczność tego rozwiązania, bowiem próby dokonywania zgłoszeń za jego pośrednictwem kończyły się komunikatem o błędzie, uniemożliwiając ich skuteczne przesłanie.

OBRAZ 2. **Usługa czatbota umożliwiająca zgłaszanie treści na platformie Telegram**



ŹRÓDŁO: TELEGRAM.COM

W ogólnodostępnym formularzu na platformie Telegram dostępne są poniższe kategorie:

a. **Nie podoba mi się**
treści, które użytkownik uważa za nieinteresujące, nieodpowiednie lub których nie chce oglądać, mimo że mogą nie naruszać zasad platformy.

b. **Przemoc wobec dzieci**
treści przedstawiające, promujące lub wspierające wykorzystywanie, krzywdzenie albo narażanie dzieci na niebezpieczeństwo.

c. **Przemoc**
materiały przedstawiające, promujące lub pochwalające przemoc fizyczną, groźby, okrucieństwo lub inne szkodliwe działania.

d. **Nielegalne towary**
sprzedaż, zakup, promocja lub dystrybucja towarów, których obrót jest zabroniony lub ograniczony przez prawo.

e. **Nielegalne treści dla dorosłych**
treści seksualne lub pornograficzne naruszające prawo, w tym materiały związane z wykorzystywaniem seksualnym, handlem ludźmi lub innymi nielegalnymi praktykami.

f. **Dane osobowe**
publikowanie lub udostępnianie prywatnych danych osobowych bez zgody osoby, której dotyczą.

g. **Terroryzm**
promowanie, wspieranie, rekrutowanie lub gloryfikowanie organizacji terrorystycznych, ideologii terrorystycznych lub aktów terroru.

h. **Oszustwo lub spam**
phishing, wyłudzanie danych lub pieniędzy, scamy, spam oraz inne działania mające na celu wprowadzenie użytkowników w błąd.

i. **Prawa autorskie**
naruszenia praw autorskich, znaków towarowych lub innych praw własności intelektualnej.

j. **Inne**
zgłoszenia dotyczące treści, które nie mieszczą się w żadnej z powyższych kategorii.

W ramach podkategorii „Nielegalne towary" dostępne są poniższe możliwości wyboru:

a. **Broń**
sprzedaż, zakup, promowanie, dystrybucja lub instrukcje dotyczące broni palnej, amunicji oraz innych rodzajów uzbrojenia.

b. **Narkotyki**
handel, produkcja, dystrybucja lub promowanie narkotyków, substancji odurzających i innych nielegalnych środków psychoaktywnych.

c. **Fałszywe dokumenty**
tworzenie, sprzedaż, zakup lub udostępnianie podrobionych dokumentów, takich jak dowody tożsamości, paszporty, prawa jazdy czy certyfikaty.

d. **Fałszywe pieniądze**
produkcja, sprzedaż, zakup lub rozpowszechnianie podrobionych banknotów, monet oraz innych środków płatniczych.

e. **Inne rzeczy**
pozostałe nielegalne towary, które nie mieszczą się w powyższych kategoriach, ale których posiadanie, sprzedaż lub dystrybucja narusza obowiązujące przepisy prawa.

W ramach podkategorii „Oszustwo lub spam" dostępne są poniższe możliwości wyboru:

a. **Phishing**
próby wyłudzenia poufnych informacji, takich jak hasła, dane logowania, dane osobowe lub dane finansowe, poprzez podszywanie się pod instytucje lub osoby powszechnie znane.

b. **Podszywanie się**
tworzenie lub wykorzystywanie fałszywej tożsamości w celu udawania innej osoby, firmy lub organizacji i wprowadzania użytkowników w błąd.

c. **Nieuczciwa sprzedaż**
oszukańcze oferty sprzedaży, wprowadzające w błąd informacje o produktach lub usługach, pobieranie płatności bez realizacji transakcji oraz inne nieuczciwe praktyki handlowe.

d. **Spam**
masowe, niechciane lub powtarzające się wiadomości, treści reklamowe albo działania mające na celu zalewanie użytkowników niepożądaną komunikacją.

W ramach podkategorii „Nielegalne treści dla dorosłych" dostępne są poniższe możliwości wyboru:

a. **Przemoc wobec dzieci**
treści przedstawiające, promujące lub wspierające wykorzystywanie, krzywdzenie lub narażanie dzieci na niebezpieczeństwo.

b. **Sex materiały wykonane bez twojej zgody**
treści o charakterze seksualnym udostępniane lub tworzone bez zgody osoby, której dotyczą, w tym intymne zdjęcia, nagrania oraz materiały typu revenge porn.

c. **Inne nielegalne treści seksualne**
treści seksualne naruszające prawo, które nie należą do powyższych kategorii, w tym materiały związane z wykorzystywaniem seksualnym, przymusem lub innymi nielegalnymi działaniami.

W ramach podkategorii „Dane osobowe" dostępne są poniższe możliwości wyboru:

a. **Prywatne zdjęcia**
publikowanie, udostępnianie lub rozpowszechnianie prywatnych zdjęć bez zgody osoby, której dotyczą.

b. **Numer telefonu**
udostępnianie, publikowanie lub rozpowszechnianie numerów telefonów bez zgody ich właścicieli.

c. **Adres**
publikowanie lub ujawnianie adresów zamieszkania, adresów korespondencyjnych lub innych danych lokalizacyjnych bez zgody zainteresowanej osoby.

d. **Inne informacje osobowe**
udostępnianie, publikowanie lub rozpowszechnianie innych danych osobowych, takich jak adresy e-mail, numery identyfikacyjne, dane finansowe lub inne informacje umożliwiające identyfikację osoby.

Discord

Zespół OAD NASK nie posiada na platformie komunikacyjnej Discord punktu kontaktowego do zgłaszania treści dezinformacyjnych. Z tego powodu zgłaszanie materiałów możliwe jest jedynie poprzez ogólny formularz w Centrum Pomocy Discord oraz z poziomu użytkownika w samej aplikacji.

Ogólny formularz zgłoszeń zawiera następujące kategorie kontaktowe:

- Pomoc i wsparcie,
- Zwroty, anulowanie i płatności,
- Zhakowane konto,
- Odwołania i prośby o aktualizację wieku,
- Zgłaszanie błędów,
- Błędy w tłumaczeniu,
- Żądanie usunięcia treści zgodnie z ustawą DMCA,
- Złożenie roszczenia wzajemnego DMCA,
- Złożenie skargi dotyczącej znaków towarowych,
- Pomoc związana z monetyzacją serwera,
- Odwołania i pytania właściciela serwera NSFW,
- Kontakt z zespołem ds. prywatności Discorda,
- Zespół wsparcia deweloperów ds. zgodności,
- Zespół wsparcia dla deweloperów.

Jeśli chodzi o zgłaszanie z poziomu użytkownika w aplikacji Discord, to dostępne są następujące kategorie:

- Nie lubię tego,
- Spam,
- Nadużycie lub nękanie,
- Szkodliwa dezinformacja lub gloryfikowanie przemocy,
- Ujawnianie prywatnych danych identyfikacyjnych,
- Coś innego,
- Zgłoś nielegalne treści na mocy aktu o usługach cyfrowych.

LinkedIn

W przypadku sieci społecznościowej LinkedIn zespół OAD NASK nie ma utworzonego punktu kontaktowego. W związku z tym jedyną opcją przekazywania treści dezinformacyjnych do platformy jest zgłaszanie ich z poziomu użytkownika.

Bluesky

Serwis Bluesky nie posiada dedykowanego narzędzia umożliwiającego zgłaszanie treści o charakterze dezinformacyjnym. Zgłoszenia mogą być przekazywane za pośrednictwem ogólnodostępnego formularza. Wypełniająca go osoba nie otrzymuje potwierdzenia przyjęcia sprawy, w tym numeru referencyjnego ani informacji o jej dalszym statusie.

W ramach formularza użytkownik lub użytkowniczka ma możliwość przedstawienia szczegółowego opisu zgłoszenia bez względu na przypisaną kategorię (nie jest to wymagane). Z racji rosnącej popularności platformy Bluesky, a wraz z tym wzrostu ilości treści dezinformacyjnych na niej, z inicjatywy OAD ustanowiono punkt kontaktowy z platformą. Umożliwi on bezpośrednie przekazywanie problematycznych spraw do administracji serwisu Bluesky.

W ramach formularza zgłoszeniowego na platformie Bluesky dostępne są poniższe kategorie:

a. **Misleading (Wprowadzanie w błąd)**
   spam lub inne nieautentyczne zachowania i metody wprowadzania użytkowników w błąd.

b. **Adult content (Treści dla dorosłych)**
   nagość, materiały dla dorosłych oraz treści seksualne publikowane bez zgody osoby, której dotyczą.

c. **Harassment or hate (Nękanie lub nienawiść)**
   obraźliwe lub dyskryminujące zachowania, nękanie, zastraszanie oraz treści szerzące nienawiść.

d. **Violence (Przemoc)**
   brutalne lub zawierające groźby treści przedstawiające, promujące lub pochwalające przemoc.

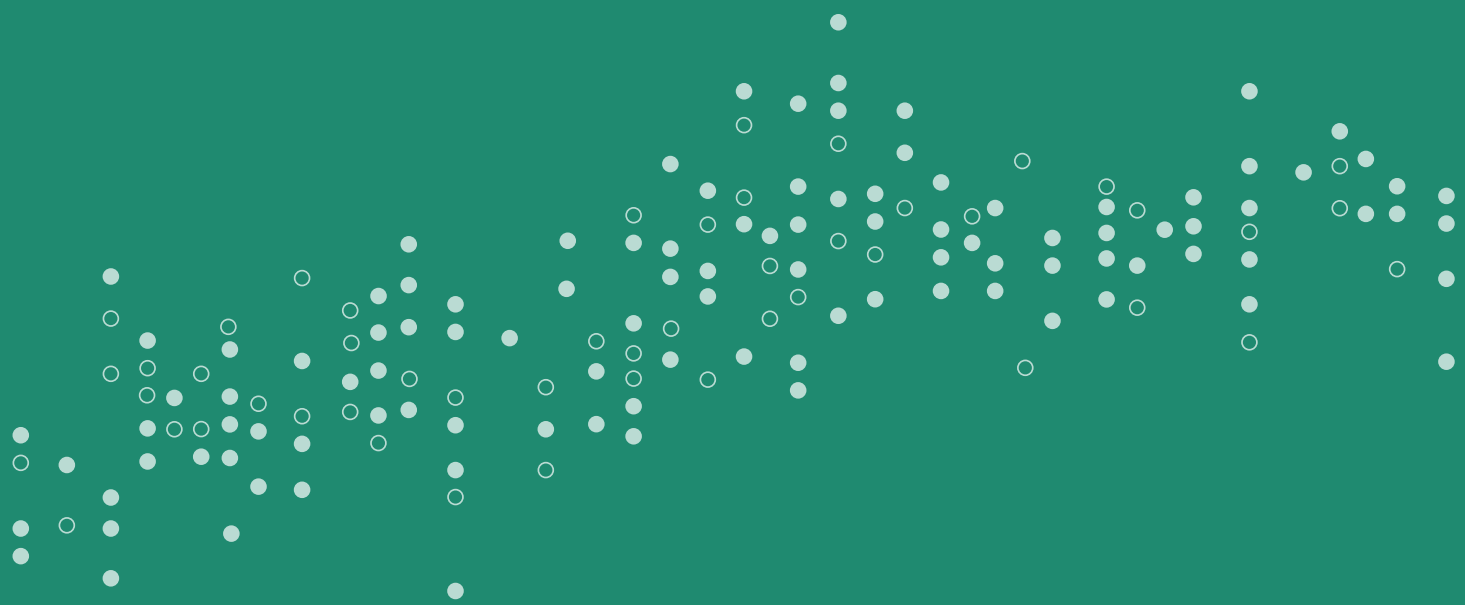
e. **Child safety (Bezpieczeństwo dzieci)**
treści związane z wykorzystywaniem, krzywdzeniem lub narażaniem osób niepełnoletnich.

f. **Self-harm or dangerous behaviors (Samookaleczenie lub niebezpieczne zachowania)**
treści promujące lub przedstawiające samookaleczenia oraz działania wysokiego ryzyka.

g. **Breaking site rules (Naruszenie zasad serwisu)**
treści objęte banem lub naruszające zasady bezpieczeństwa platformy.

h. **Inne**
zgłoszenia dotyczące treści, które nie mieszczą się w żadnej z powyższych kategorii.

Kluczowe wnioski



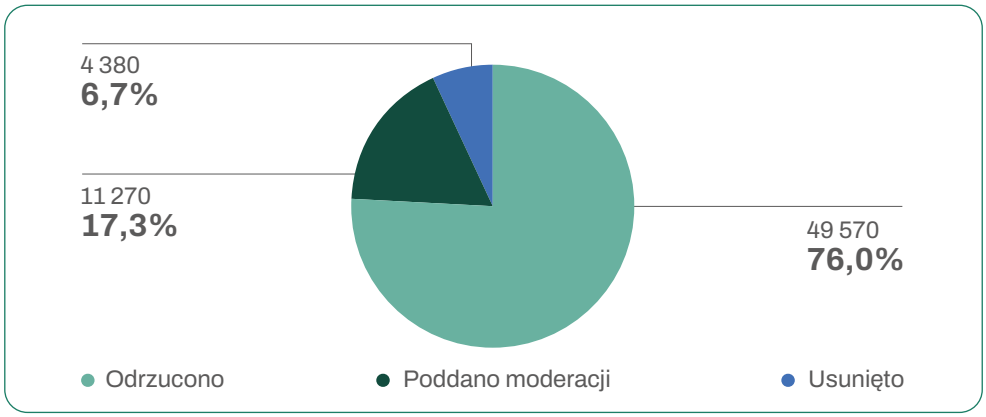
Kluczowe wnioski

W pierwszym półroczu 2026 roku zgłoszono 65 220 treści dezinformacyjnych. Platformy usunęły 4380 (6,7 proc.) z nich, 11 270 (17,3 proc.) poddały moderacji, a 49 570 materiałów (76 proc.) pozostawiły bez żadnej reakcji.

Główne wnioski są następujące:

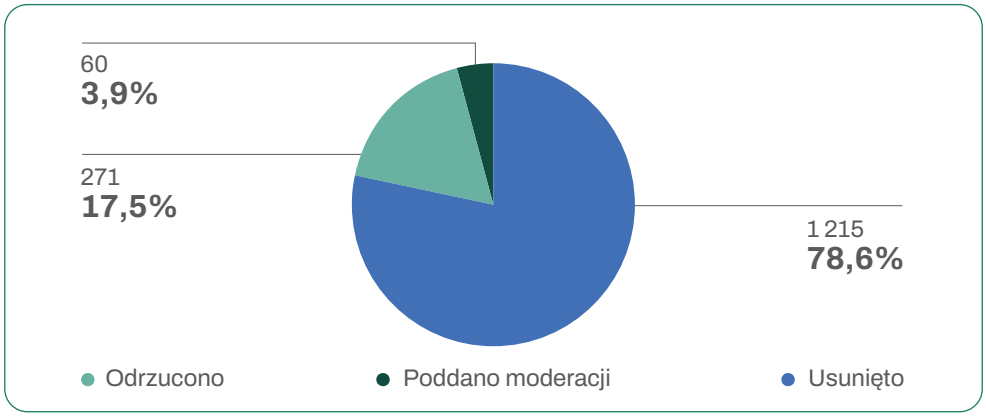
- Spośród 65 220 zgłoszeń do kategorii fałszywa informacja przypisano 1537 materiałów (2,3 proc.) i stanowią one najmniej liczną kategorię w opisywanym okresie. 62 137 materiałów (95,3 proc.) zawierało manipulacje. Natomiast 1546 incydentów (2,4 proc.) związanych było z podszywaniem się pod instytucję lub osobę powszechnie znaną.
- Platformy najczęściej podejmowały działania w odpowiedzi na incydenty związane z podszywaniem się pod instytucję lub osobę powszechnie znaną. W analizowanym okresie wśród zgłoszeń z tej kategorii 1215 materiałów (78,6 proc.) zostało usuniętych; 60 incydentów (3,9 proc.) zostało poddanych moderacji; 271 treści (17,5 proc.) nie doczekało się reakcji i pozostaje dostępnych.
- W odpowiedzi na zgłoszenia dotyczące zmanipulowanych treści platformy społecznościowe zareagowały w następujący sposób: 3021 materiałów (4,9 proc.) zostało usuniętych; 11 125 treści (17,9 proc.) zostało poddanych moderacji; 47 991 incydentów (77,2 proc.) nie spotkało się z żadną reakcją zespołów moderujących. Przeważająca liczba zgłoszeń z tej kategorii dotyczyła kont zakwalifikowanych na podstawie wewnętrznej analizy OAD NASK jako boty działające w ramach CIB.
- Reakcja platform w zakresie moderacji treści przypisanych do kategorii fałszywa informacja pozostaje najbardziej ograniczona. Spośród wszystkich 1537 zgłoszeń usunięto 144 treści (9,4 proc.), natomiast moderacji poddano 85 materiałów (5,5 proc.). Pozostałych 1308 zgłoszeń (85,1 proc.) zostało odrzuconych.

WYKRES 1. Reakcja platform na wszystkie zgłoszenia



ŹRÓDŁO: OPRACOWANIE WŁASNE

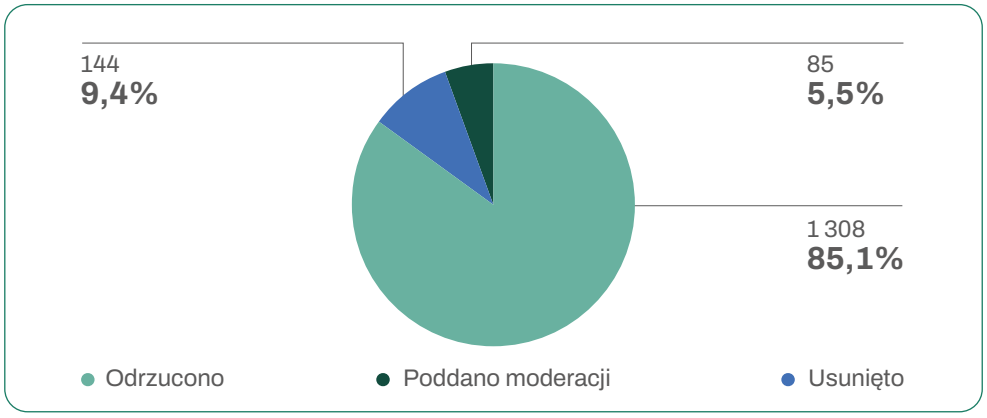
WYKRES 2. Reakcja platform na zgłoszenia podszywania



ŹRÓDŁO: OPRACOWANIE WŁASNE

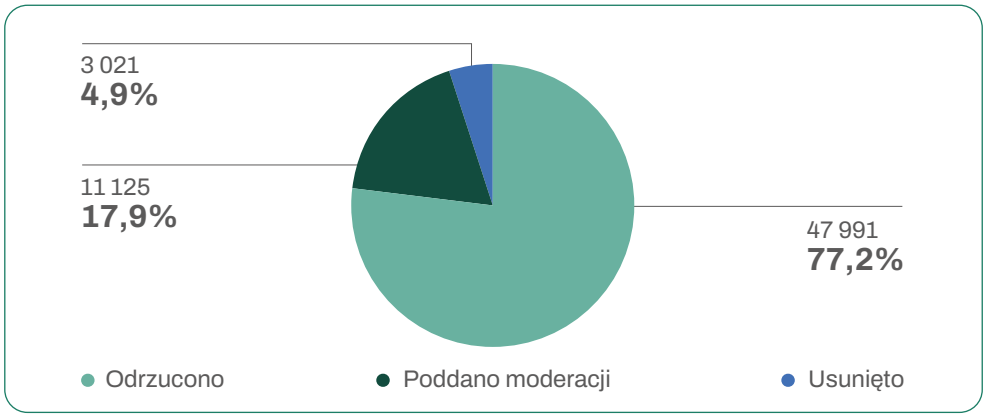
Kluczowe wnioski

WYKRES 3. Reakcja platform na zgłoszenia fałszywej informacji



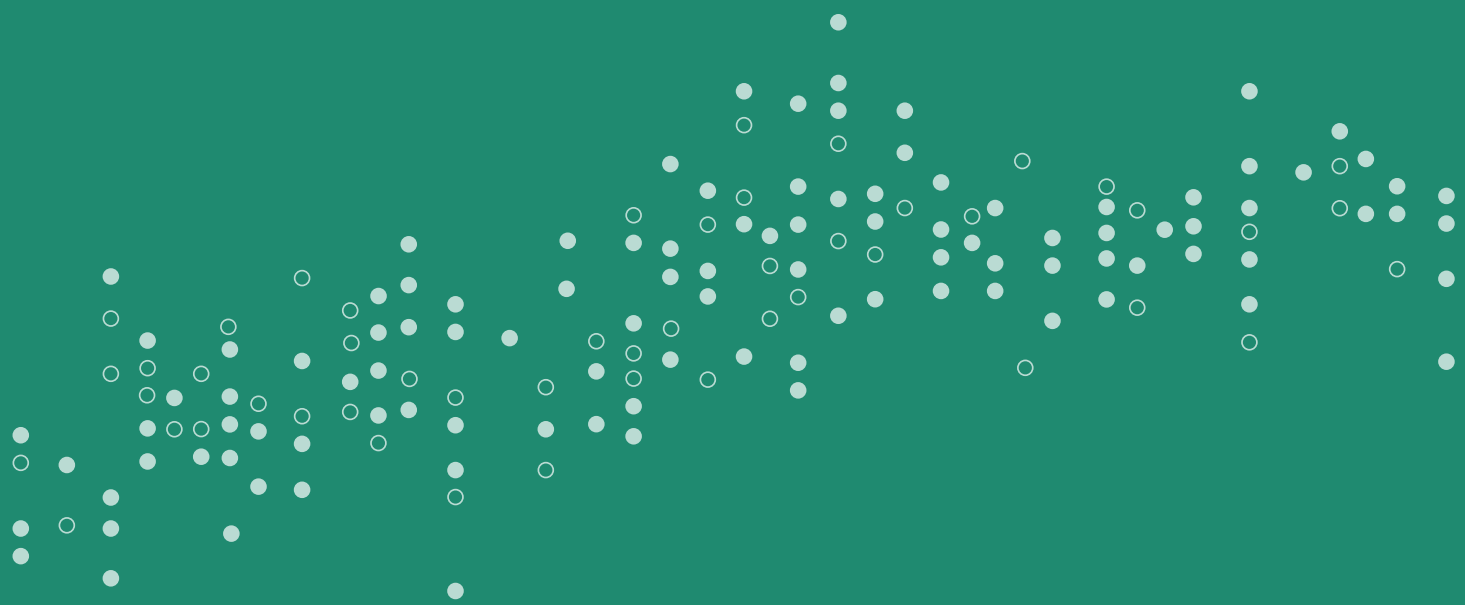
ŹRÓDŁO: OPRACOWANIE WŁASNE

WYKRES 4. Reakcja platform na zgłoszenia manipulacji



ŹRÓDŁO: OPRACOWANIE WŁASNE

Podsumowanie pierwszego półrocza 2026 roku w podziale na platformy społecznościowe



Podsumowanie pierwszego półrocza 2026 roku w podziale na platformy społecznościowe

Na wykresie 5. przedstawiono reakcje platform społecznościowych na incydenty zgłoszone przez analityków i analityczki OAD NASK. W I półroczu 2026 roku materiały przekazano między innymi do serwisów Facebook, Instagram, Threads, X, TikTok, Google, Discord, Telegram oraz Bluesky. Wykres uwzględnia również reakcje serwisów oznaczonych jako „Inne". W tej grupie znalazł się między innymi serwis wykop.pl.

W analizowanym okresie administracja platformy X podjęła decyzję o usunięciu 1438 publikacji, co odpowiada 2,4 proc. zgłoszonych treści. Zespół moderacji zdecydował o odrzuceniu 78,9 proc. zgłoszeń, co oznacza, że użytkownicy i użytkowniczki nadal mają dostęp do 47 464 materiałów zawierających dezinformację. Moderacji poddano natomiast 11 230 treści, a więc 18,7 proc. zgłoszeń.

Platforma Facebook usunęła 2150 materiałów (63,6 proc. zgłoszeń). Administracja odrzuciła 1194 zgłoszeń (35,3 proc. przypadków), a 37 incydentów (1,1 proc. zgłoszeń) zostało poddanych moderacji. Platforma Instagram usunęła 59 materiałów, co odpowiada 37,8 proc. zgłoszonych treści. Odrzucono 97 zgłoszeń, czyli 62,2 proc. przypadków. Serwis Threads usunął 2 materiały, co odpowiada 6,9 proc. zgłoszonych treści, odrzucono natomiast 27 zgłoszeń, czyli 93,1 proc. przypadków.

Administracja platformy TikTok usunęła 618 zgłoszonych treści, co stanowi 55,7 proc. przypadków. 488 materiałów, czyli 44 proc. zgłoszeń, odrzucono. Moderacji poddano 3 materiały, a więc 0,3 proc. treści.

Platformy Google, obejmujące takie serwisy i usługi jak Google Ads oraz YouTube, usunęły łącznie 42 materiałów, co stanowi 13 proc. zgłoszeń. Brak reakcji stwierdzono w przypadku 285 zgłoszeń, czyli 87 proc. treści.

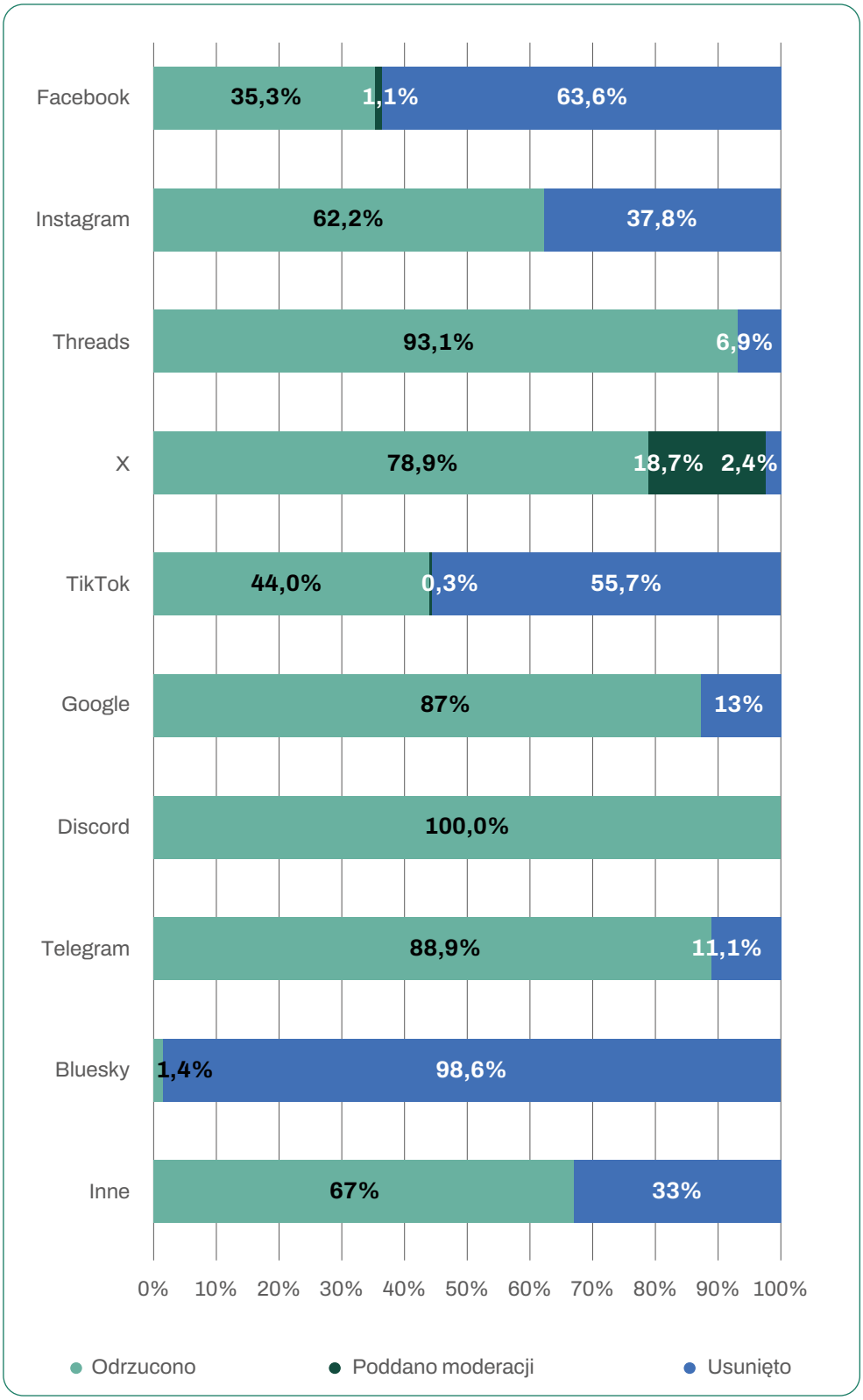
Administratorzy serwisu Discord odrzucili wszystkie 4 zgłoszenia.

Platforma Telegram podjęła reakcję w przypadku 1 materiału, co przekłada się na 11,1 proc. zgłoszeń, a odrzuciła pozostałych 8 zgłoszeń, czyli 88,9 proc. zaraportowanych incydentów.

Platforma Bluesky usunęła 69 incydentów, co stanowi 98,6 proc. zgłoszonych treści, a odrzuciła 1 zgłoszenie, co odpowiada 1,4 proc. materiałów.

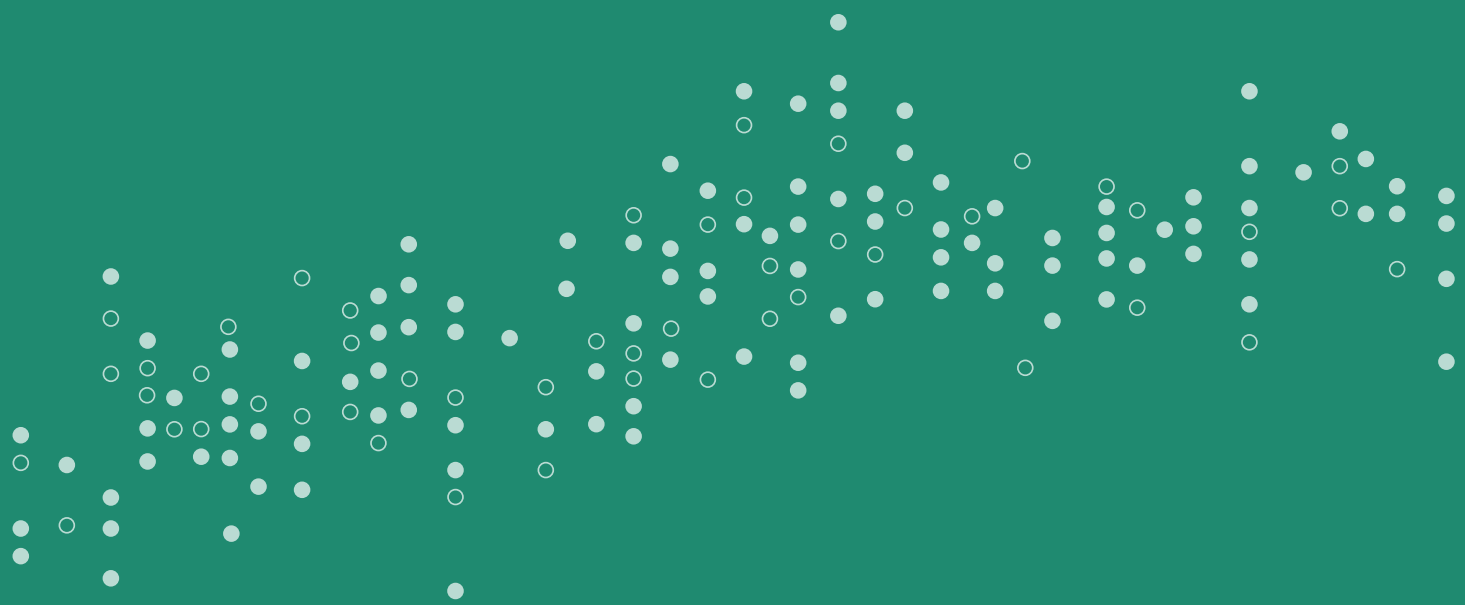
Osoby odpowiedzialne za administrację pozostałych platform usunęły 1 materiał, a pozostałe 2 zgłoszenia odrzuciły.

WYKRES 5. Reakcja platform społecznościowych na zgłoszone incydenty



ŹRÓDŁO: OPRACOWANIE WŁASNE

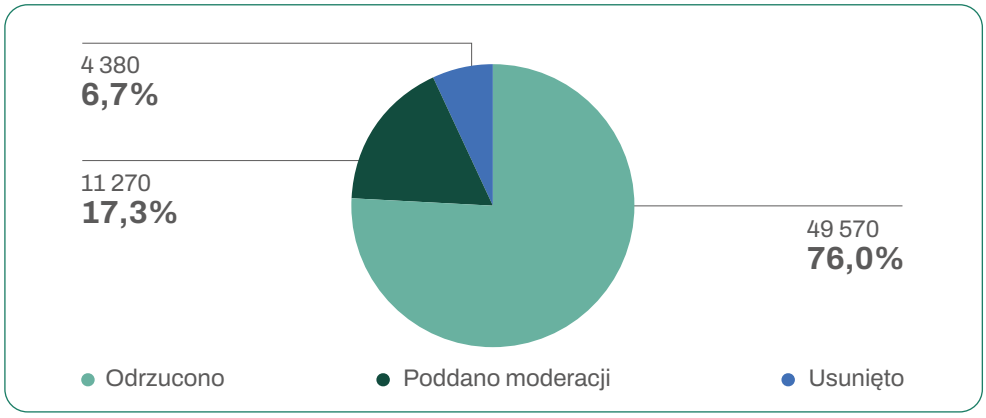
Reakcja platform na wszystkie zgłoszenia OAD NASK



Reakcja platform na wszystkie zgłoszenia OAD NASK

Z platform społecznościowych usunięto 4380 publikacji (6,7 proc.). Moderacji zostało poddanych 11 270 wpisów (17,3 proc.), a 49 570 zgłoszonych treści (76 proc.) nie spotkało się z żadną widoczną reakcją ze strony zespołów moderujących.

WYKRES 6. **Reakcja platform na wszystkie zgłoszenia**



ŹRÓDŁO: OPRACOWANIE WŁASNE

Tabela 1 przedstawia zgłoszenia według wewnętrznej kategoryzacji z podziałem na rodzaj naruszenia. Wśród wszystkich zgłoszeń zidentyfikowano:

- 62 137 wpisów (95,3 proc.) zawierających zmanipulowane informacje,
- 1546 incydenty (2,4 proc.) związane z podszywaniem się pod osobę lub instytucję,
- 1537 materiałów (2,3 proc.) zawierających fałszywe informacje.

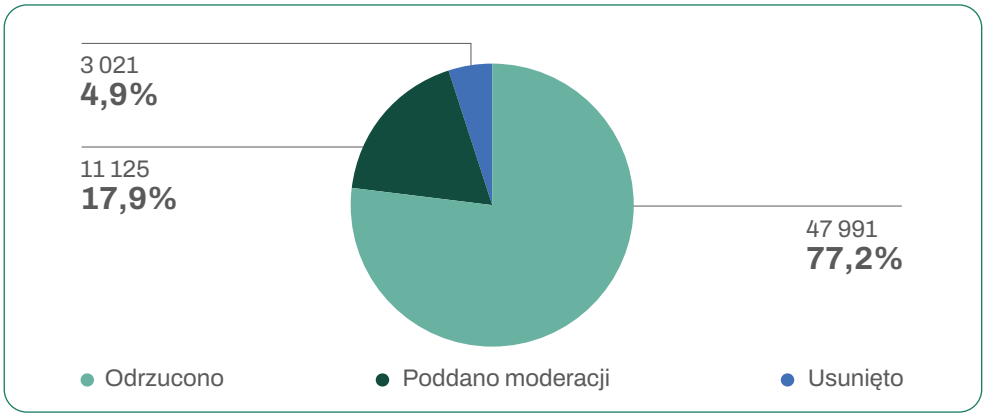
TABELA 1. Rodzaj przypisanego naruszenia dla wszystkich zgłoszeń

Rodzaj naruszenia	Fałszywa informacja	Manipulacja	Podszywanie się pod osobę/instytucję
Liczba zgłoszeń odrzuconych	1 308	47 991	271
Liczba zgłoszeń poddanych moderacji	85	11 125	60
Liczba zgłoszeń usuniętych	144	3 021	1 215
Całkowita liczba zgłoszeń	1 537	62 137	1 546

ŹRÓDŁO: OPRACOWANIE WŁASNE

Wśród zgłoszeń przekazywanych przez zespół OAD NASK do platform społecznościowych najwięcej incydentów polegało na rozpowszechnianiu zmanipulowanych treści. Spośród wszystkich zgłoszonych przypadków 3021 publikacji (4,9 proc.) zostało usuniętych, a 11 125 wpisów (17,9 proc.) zostało poddanych moderacji, ale 47 991 przypadków (77,2 proc.) zmanipulowanych treści pozostaje dostępnych. Jako incydenty wystąpienia manipulacji zakwalifikowane zostały między innymi treści publikowane przez konta-boty, czyli profile wykazujące oznaki zautomatyzowanego działania.

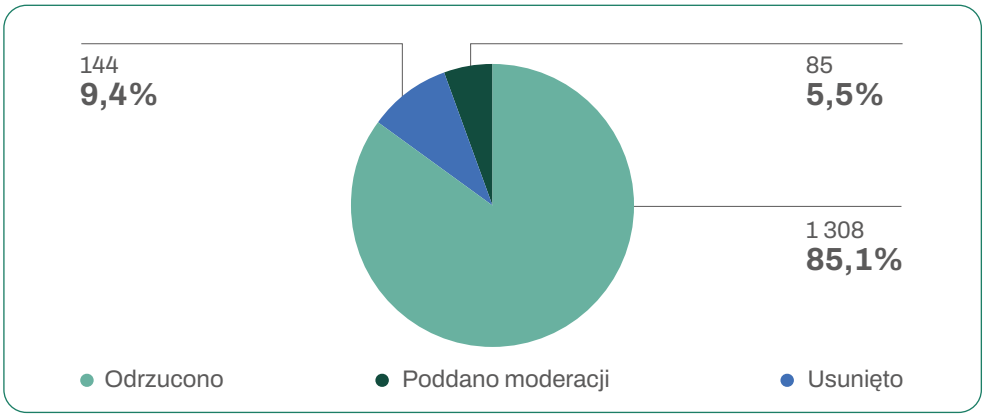
WYKRES 7. Reakcja platform na zgłoszenia manipulacji



ŹRÓDŁO: OPRACOWANIE WŁASNE

W odpowiedzi na wskazania OAD NASK serwisy społecznościowe usunęły 144 materiały (9,4 proc.) zawierające fałszywe informacje. Odrzuciły 1308 zgłoszeń (85,1 proc.) z tej kategorii, a 85 materiałów (5,5 proc.) poddały moderacji.

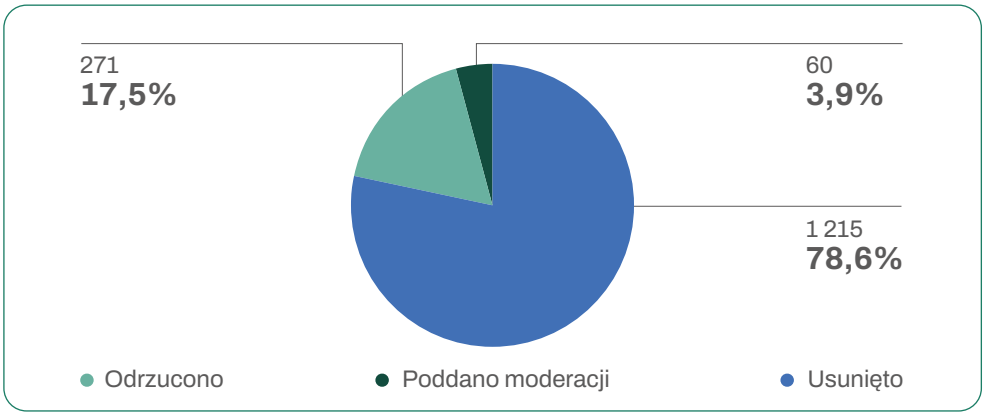
WYKRES 8. **Reakcja platform na zgłoszenia fałszywej informacji**



ŹRÓDŁO: OPRACOWANIE WŁASNE

Jeśli chodzi o zidentyfikowane przez Ośrodek przypadki podszywania się pod osoby powszechnie znane lub instytucje, usunięto 1215 materiałów (78,6 proc.). Dalszych 60 incydentów (3,9 proc.) zostało poddanych moderacji, a 271 zgłoszeń (17,5 proc.) odrzucono.

WYKRES 9. **Reakcja platform na zgłoszenia podszywania**

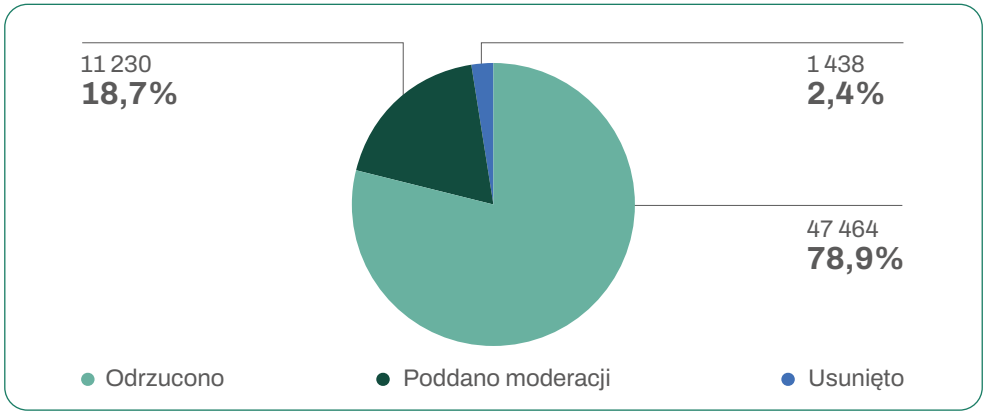


ŹRÓDŁO: OPRACOWANIE WŁASNE

X.com

Platforma X podjęła decyzję o usunięciu 1438 publikacji, co odpowiada 2,4 proc. zgłoszonych treści. Administracja serwisu zdecydowała o odrzuceniu 78,9 proc. zgłoszeń, co oznacza, że użytkownicy i użytkowniczki nadal mają dostęp do 47 464 materiałów zawierających dezinformację. Moderacji poddano natomiast 11 230 treści, a więc 18,7 proc. zgłoszeń.

WYKRES 10. **Reakcja platformy X na wszystkie zgłoszenia**

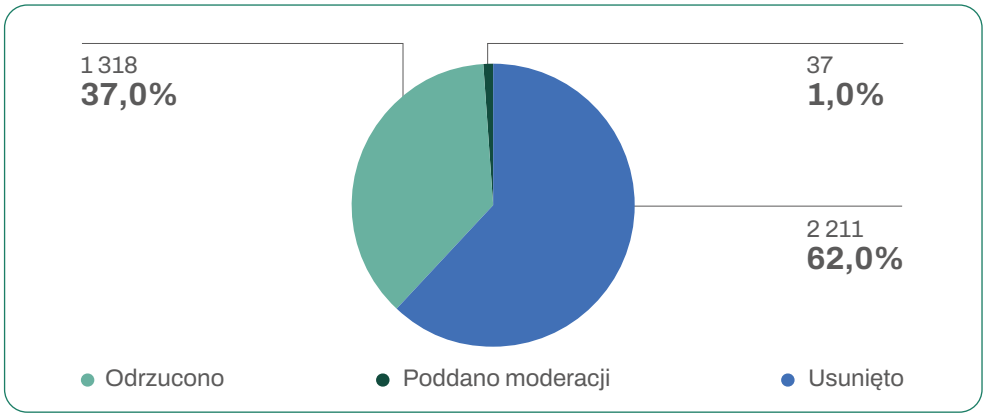


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Meta (Facebook, Instagram, Threads)

Wykres przedstawia dane dotyczące treści zgłoszonych do platform Meta, czyli serwisów Facebook, Instagram oraz Threads. Spośród wszystkich zgłoszeń 2211 przypadków, czyli 62 proc. zgłoszonych materiałów, zostało usuniętych przez administrację serwisu. Dalszych 37 incydentów, co odpowiada 1% zgłoszeń, poddano moderacji. Odrzucono 1318 zgłoszeń, czyli 37 proc. przypadków.

WYKRES 11. **Reakcja platform Meta na wszystkie zgłoszenia**

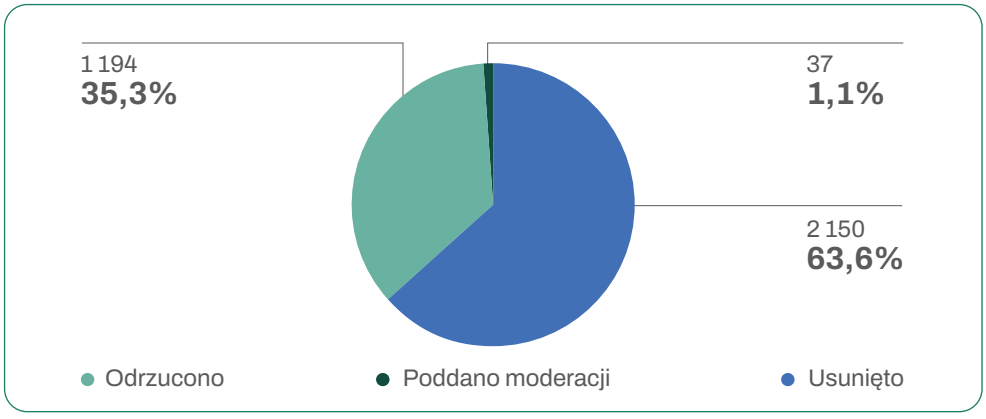


ŹRÓDŁO: OPRACOWANIE WŁASNE

Facebook

W analizowanym okresie platforma Facebook usunęła 2150 zgłoszonych materiałów (63,6 proc. treści). Administracja odrzuciła 1194 zgłoszeń (35,3 proc. przypadków), a 37 incydentów (1,1 proc. zgłoszeń) poddała moderacji.

WYKRES 12. **Reakcja platformy Facebook na wszystkie zgłoszenia**

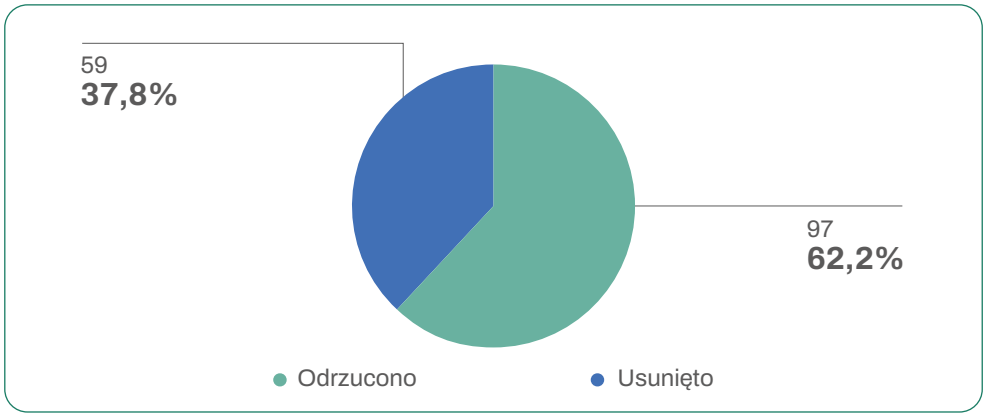


ŹRÓDŁO: OPRACOWANIE WŁASNE

Instagram

W badanym okresie platforma Instagram usunęła 59 materiałów, co odpowiada 37,8 proc. zgłoszonych treści. Odrzucono 97 zgłoszeń, czyli 62,2 proc. przypadków.

WYKRES 13. **Reakcja platformy Instagram na wszystkie zgłoszenia**

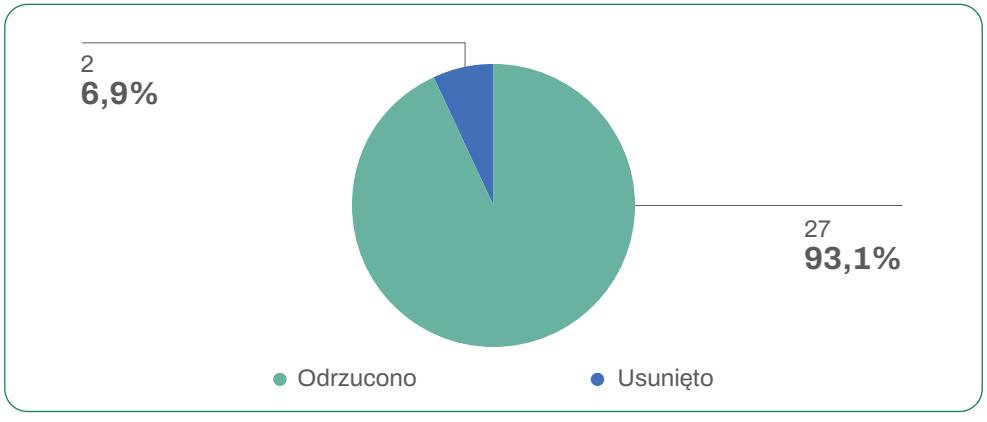


ŹRÓDŁO: OPRACOWANIE WŁASNE

Threads

W badanym okresie platforma Threads usunęła 2 materiały, co odpowiada 6,9 proc. zgłoszonych treści. Odrzucono 27 zgłoszeń, czyli 93,1 proc. przypadków.

WYKRES 14. **Reakcja platformy Threads na wszystkie zgłoszenia**

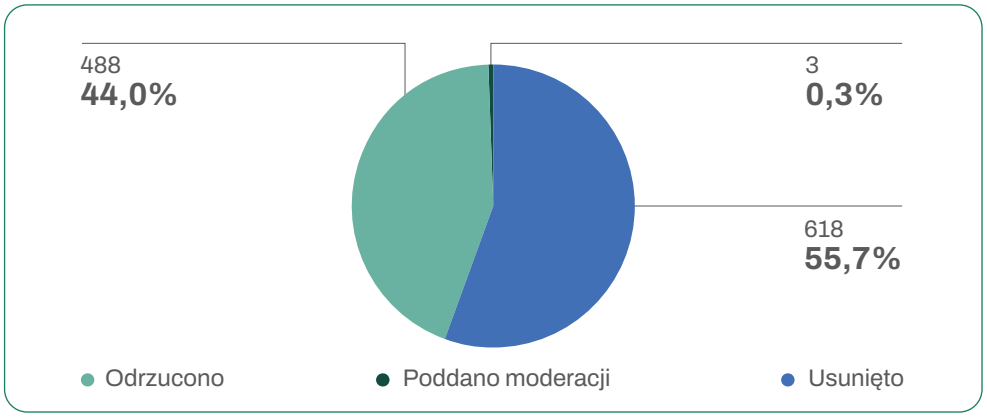


ŹRÓDŁO: OPRACOWANIE WŁASNE

TikTok

Administracja platformy TikTok usunęła 618 materiałów, co stanowi 55,7 proc. zgłoszonych treści. Z kolei 488 zgłoszeń, czyli 44 proc. przypadków, odrzucono. Moderacji poddano 3 treści, a więc 0,3 proc. materiałów.

WYKRES 15. **Reakcja platformy TikTok na wszystkie zgłoszenia**

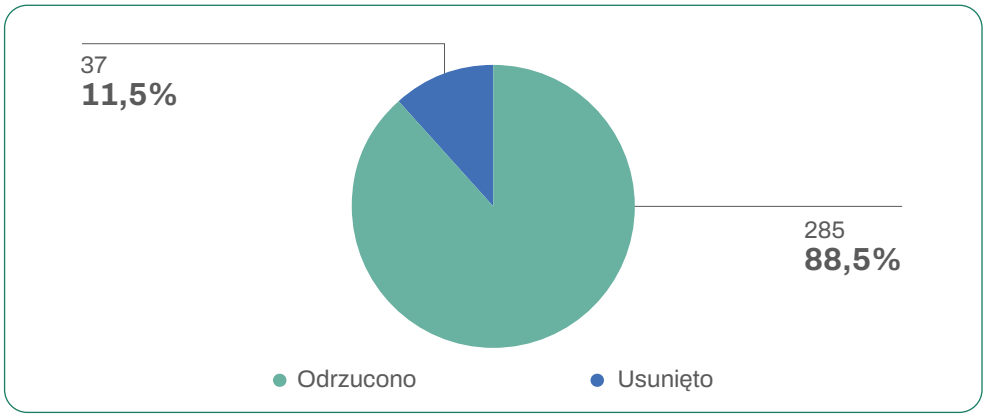


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Google (YouTube, Google Ads)

Poniżej przedstawiono łączne dane dotyczące zgłoszeń przekazanych do administracji platform Google, obejmujących takie serwisy i usługi jak Google Ads oraz YouTube. Usunięto łącznie 37 materiałów, co stanowi 11,5 proc. zgłoszeń. Reakcja nie miała miejsca w przypadku 285 treści, czyli 88,5 proc. zgłoszeń.

WYKRES 16. **Reakcja platform Google na wszystkie zgłoszenia**

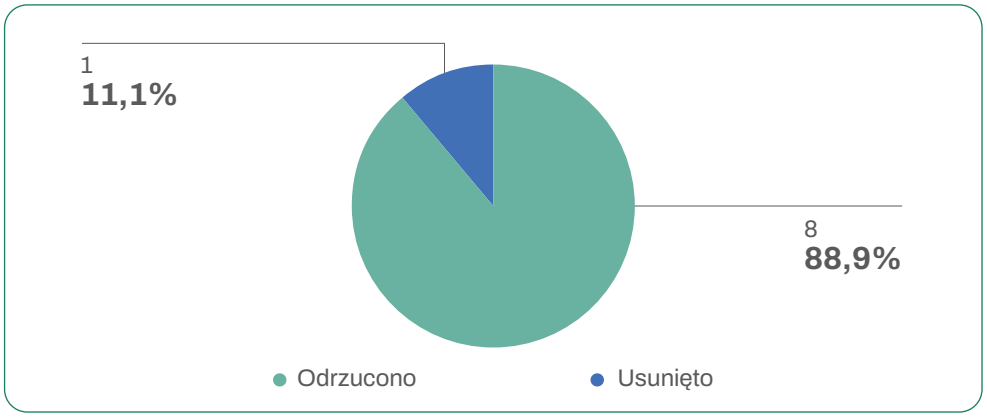


ŹRÓDŁO: OPRACOWANIE WŁASNE

Telegram

Platforma Telegram podjęła reakcję w przypadku 1 materiału, co stanowi 11,1 proc. zgłoszeń, a odrzuciła pozostałych 8 zgłoszeń, co przekłada się na 88,9 proc. zaraportowanych incydentów.

WYKRES 17. **Reakcja platformy Telegram na wszystkie zgłoszenia**



ŹRÓDŁO: OPRACOWANIE WŁASNE

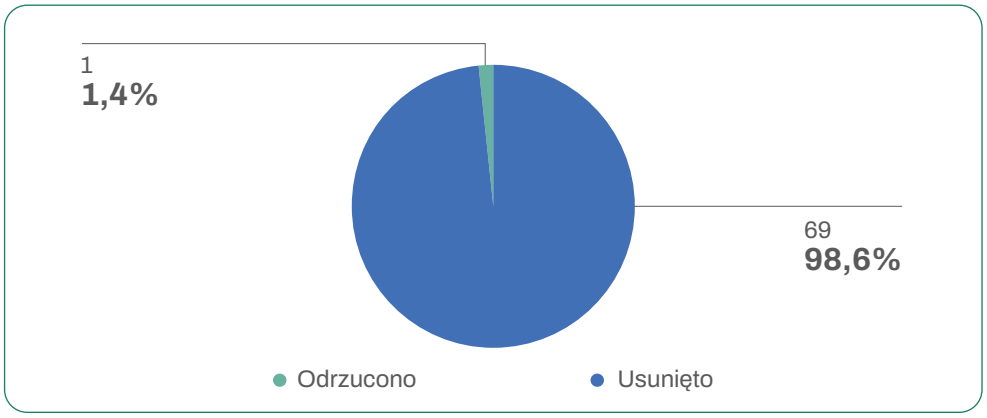
Discord

Administracja serwisu Discord odrzuciła wszystkie 4 zgłoszenia

Bluesky

Platforma Bluesky usunęła 69 wpisów, co stanowi 98,6 proc. incydentów, a odrzuciła 1 zgłoszenie, przekładające się na 1,4 proc. przypadków.

WYKRES 18. **Reakcja platformy Bluesky na wszystkie zgłoszenia**

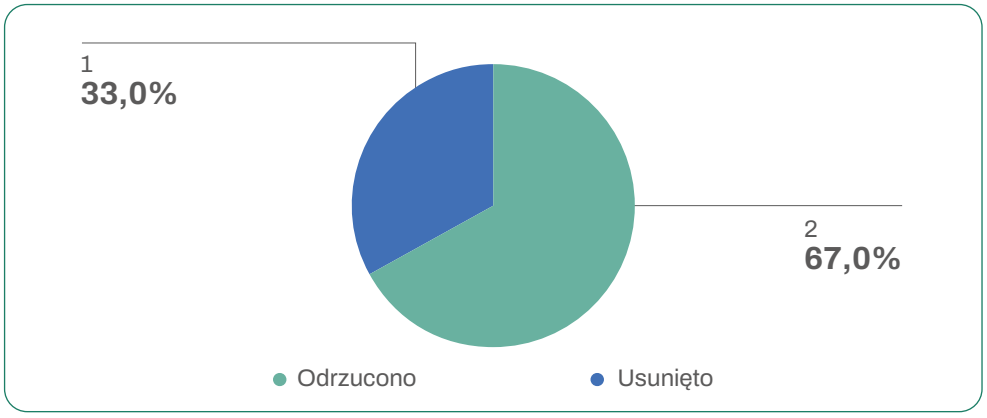


ŹRÓDŁO: OPRACOWANIE WŁASNE

Inne

Zespoły moderujące pozostałych platform usunęły 1 treść. Pozostałe 2 zgłoszenia zostały odrzucone.[2]

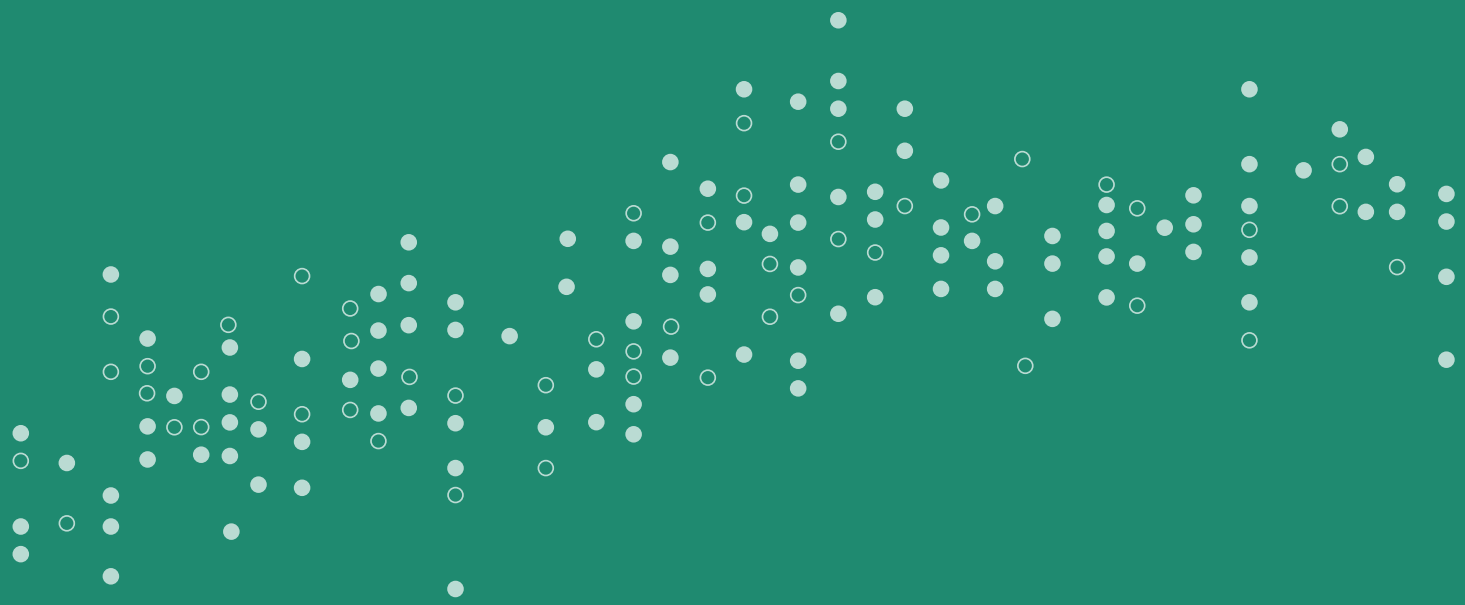
WYKRES 19. **Reakcja innych platform na wszystkie zgłoszenia**



ŹRÓDŁO: OPRACOWANIE WŁASNE

2 Platformy: Wykop, Steam

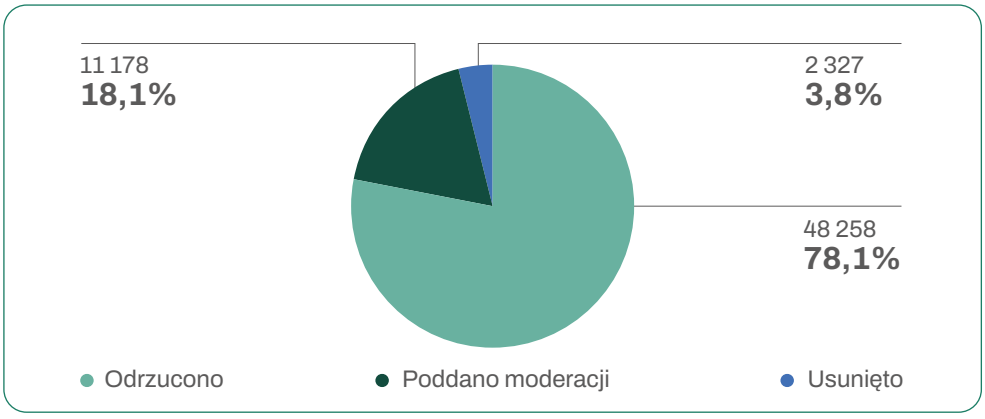
Zgłoszenia wewnętrzne



Zgłoszenia wewnętrzne

Spośród 61 763 zgłoszeń wewnętrznych administracje platform społecznościowych poddały moderacji 11 178 incydentów, co stanowi 18,1 proc. zgłoszeń. Dalszych 2327 materiałów, czyli 3,8 proc. zgłoszonych publikacji, usunięto. W stosunku do 48 258 treści, co odpowiada 78,1 proc. zgłoszeń, platformy nie podjęły żadnych działań.

WYKRES 20. **Reakcja platform na zgłoszenia wewnętrzne**



ŹRÓDŁO: OPRACOWANIE WŁASNE

Tabela 2 przedstawia zgłoszenia według wewnętrznej kategoryzacji z podziałem na rodzaj naruszenia. Wśród wszystkich zgłoszeń zidentyfikowano:

- 60 480 wpisów (98 proc.) zawierających zmanipulowane informacje,
- 1199 materiałów (1,9 proc.) zawierających fałszywe informacje,
- 84 materiały (0,1 proc.) związane z podszywaniem się pod osobę lub instytucję.

TABELA 2. Rodzaj przypisanego naruszenia dla zgłoszeń wewnętrznych

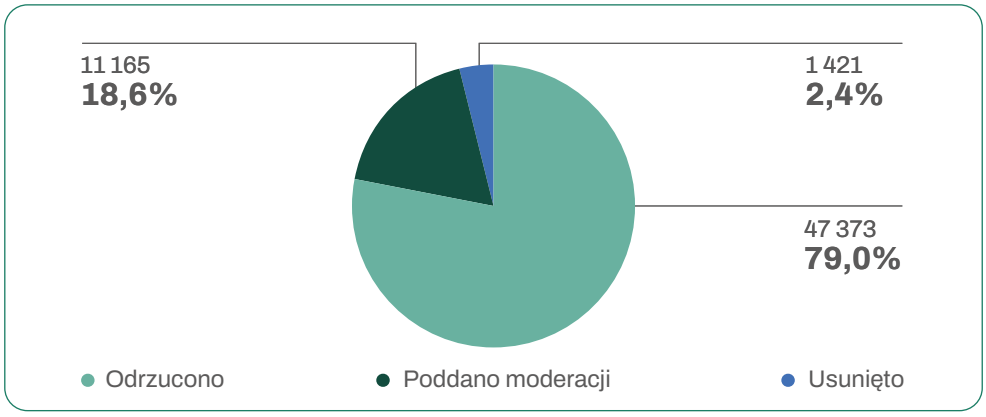
Rodzaj naruszenia	Fałszywa informacja	Manipulacja	Podszywanie się pod osobę/instytucję
Liczba zgłoszeń odrzuconych	1 071	47 186	1
Liczba zgłoszeń poddanych moderacji	73	11 101	4
Liczba zgłoszeń usuniętych	55	2 193	79
Całkowita liczba zgłoszeń	1 199	60 480	84

ŹRÓDŁO: OPRACOWANIE WŁASNE

X.com

Platforma X nie podjęła działania w przypadku 47 373 zgłoszonych treści, co odpowiada 79 proc. incydentów. Moderacji poddano 11 165 materiałów, a więc 18,6 proc. zgłoszeń. Usunięto 1421 treści, czyli 2,4 proc. incydentów.

WYKRES 21. Reakcja platformy X na zgłoszenia wewnętrzne

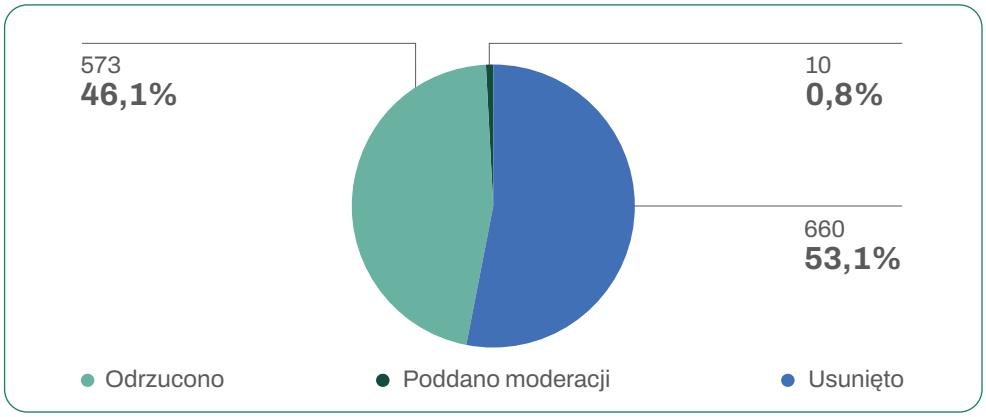


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Meta (Facebook, Instagram, Threads)

Zespoły moderujące serwisów firmy Meta usunęły 660 zgłoszonych treści, czyli 53,1 proc. incydentów. Odrzucono 573 zgłoszeń, co odpowiada 46,1 proc. materiałów. Poddane moderacji zostało 10 incydentów, czyli 0,8 proc. treści.

WYKRES 22. **Reakcja platformy Meta na wszystkie zgłoszenia wewnętrzne**

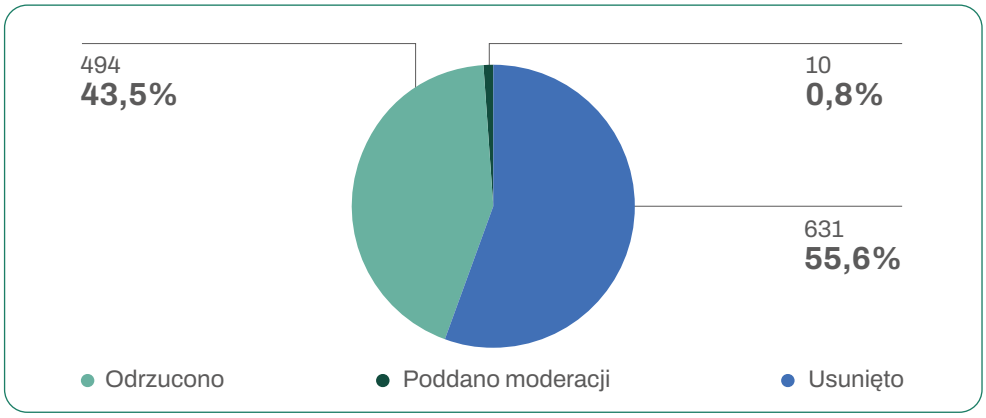


ŹRÓDŁO: OPRACOWANIE WŁASNE

Facebook

Platforma Facebook odrzuciła 43,5 proc. zgłoszeń, czyli 494 materiały. Usuniętych zostało 631 materiałów, co odpowiada 55,6 proc. przypadków. Z kolei moderacji poddano 10 incydentów dezinformacji, a więc 0,9 proc. zgłoszeń.

WYKRES 23. **Reakcja platformy Facebook na zgłoszenia wewnętrzne**

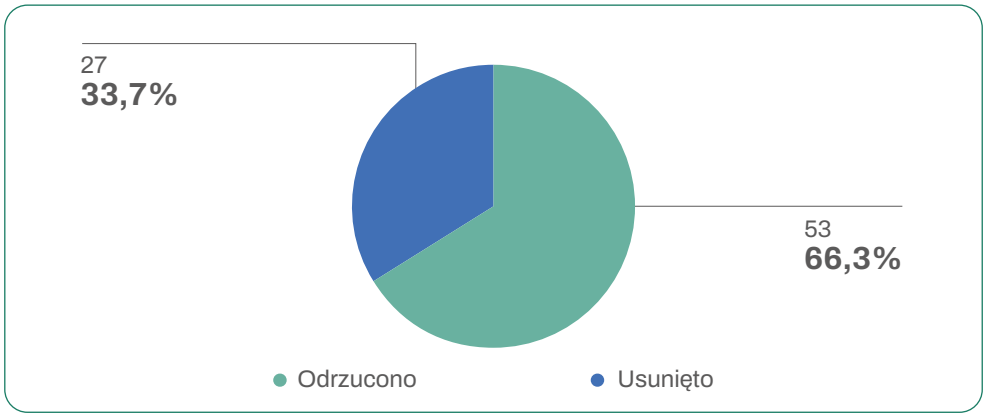


ŹRÓDŁO: OPRACOWANIE WŁASNE

Instagram

Administracja platformy Instagram usunęła 27 materiałów, a więc 33,7 proc. incydentów. Odrzuciła 53 zgłoszenia, czyli 66,3 proc. treści.

WYKRES 24. **Reakcja platformy Instagram na zgłoszenia wewnętrzne**

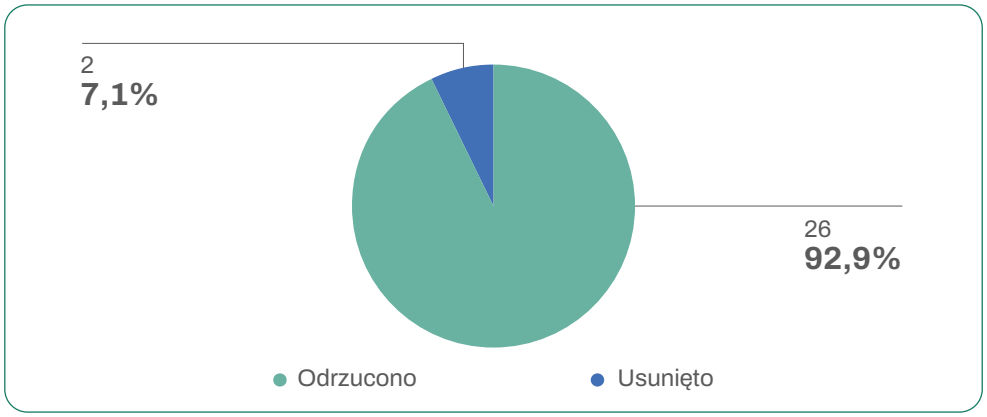


ŹRÓDŁO: OPRACOWANIE WŁASNE

Threads

Administracja platformy Threads usunęła 2 materiały, a więc 7,1 proc. incydentów. Odrzuciła z kolei 26 zgłoszeń, czyli 92,9 proc. incydentów.

WYKRES 25. **Reakcja platformy Threads na zgłoszenia wewnętrzne**

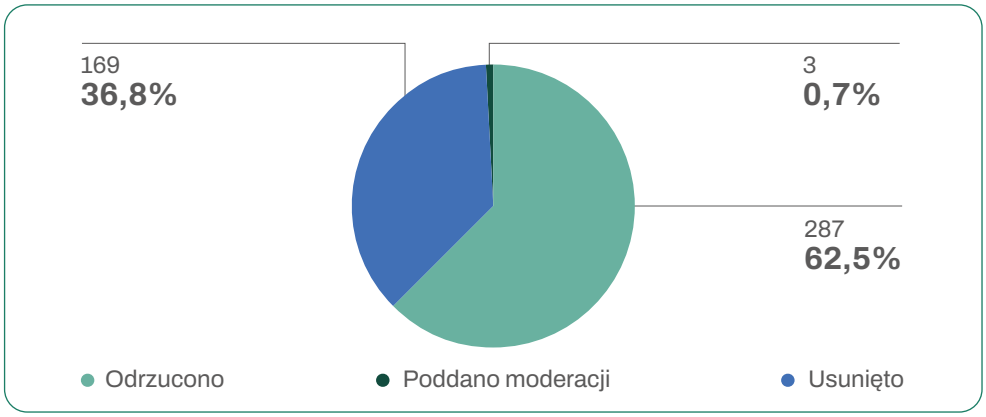


ŹRÓDŁO: OPRACOWANIE WŁASNE

TikTok

Administracja platformy TikTok usunęła 169 materiałów, co stanowi 36,8 proc. incydentów. Odrzucono 287 zgłoszeń, co odpowiada 62,5 proc. treści. Moderacji zostały poddane 3 publikacje, a więc 0,7 proc. incydentów.

WYKRES 26. **Reakcja platformy TikTok na zgłoszenia wewnętrzne**

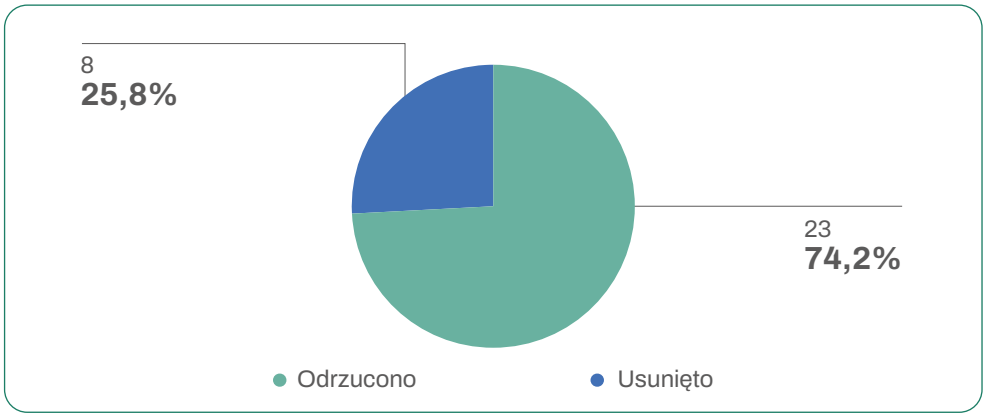


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Google (YouTube, Google Ads)

Administracja platform Google usunęła 8 treści, czyli 25,8 proc. incydentów, natomiast 23 zgłoszenia, a więc 74,2 proc. materiałów, zostały odrzucone.

WYKRES 27. **Reakcja platformy Google na zgłoszenia wewnętrzne**

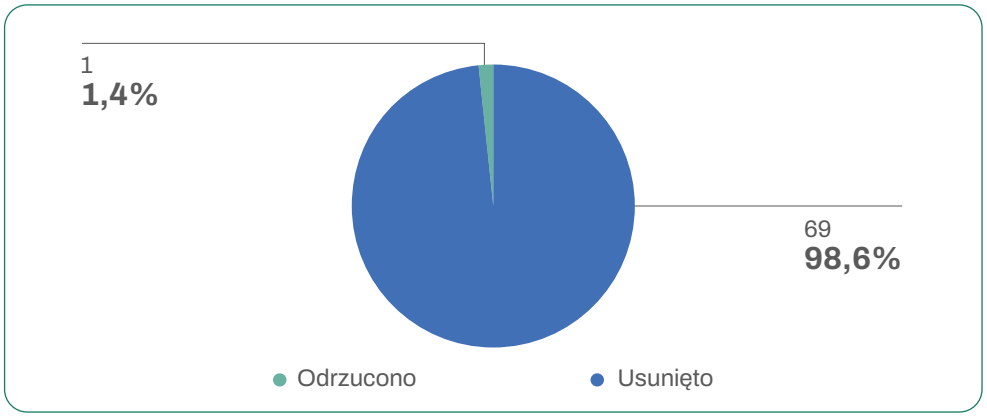


ŹRÓDŁO: OPRACOWANIE WŁASNE

Bluesky

Administratorzy serwisu Bluesky usunęli 69 materiałów, czyli 98,6 proc. zgłoszonych treści. Odrzucone zostało zgłoszenie 1 materiału, co przekłada się na 1,4 proc. treści.

WYKRES 28. **Reakcja platformy Bluesky na zgłoszenia wewnętrzne**

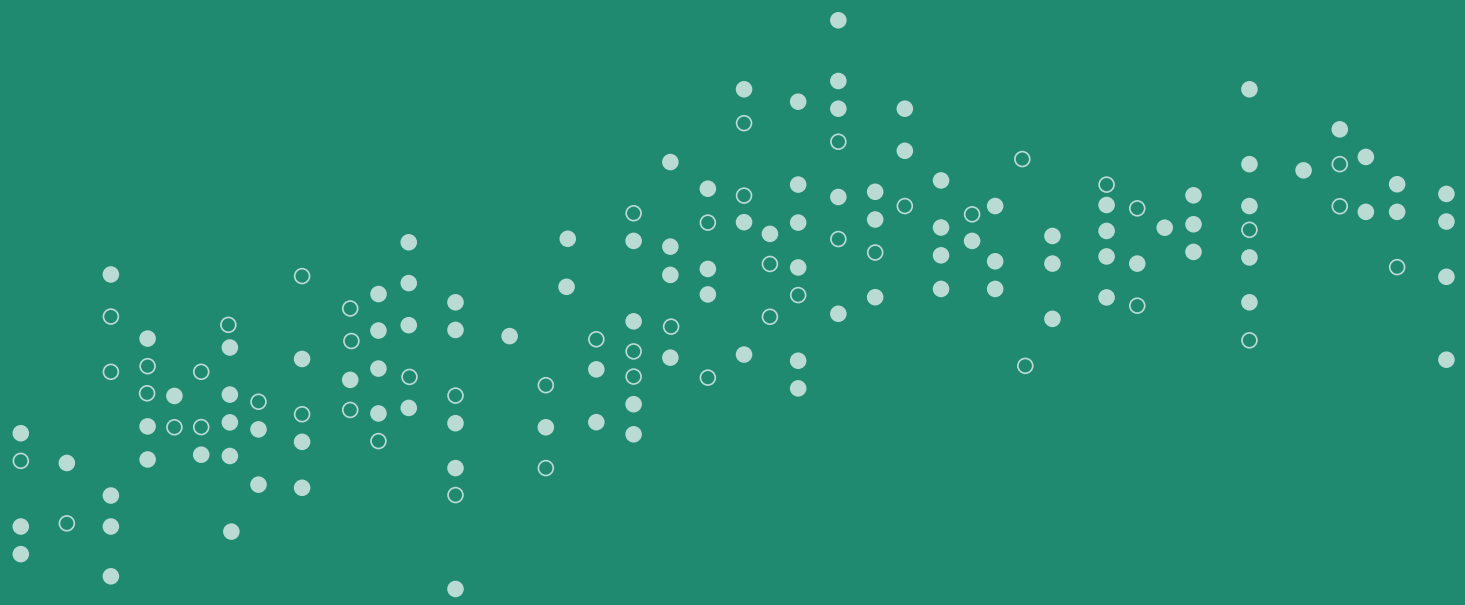


ŹRÓDŁO: OPRACOWANIE WŁASNE

Inne

Administracja platformy Steam odrzuciła 1 zgłoszenie, które zostało do nich skierowane.

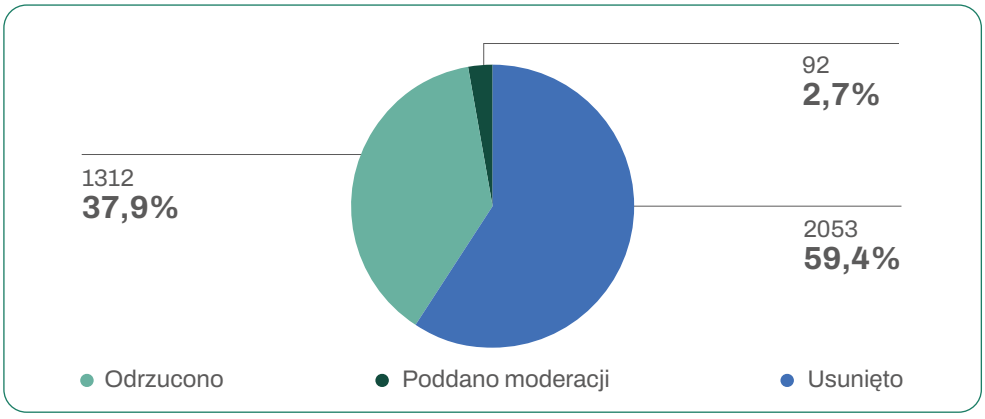
Zgłoszenia zewnętrzne



Zgłoszenia zewnętrzne

W przypadku 3457 zgłoszeń zewnętrznych, czyli tych przekazywanych do OAD NASK przez obywateli oraz instytucje, platformy usunęły 2053 incydentów, co odpowiada 59,4 proc. zgłoszeń. Kolejne 92 treści, czyli 2,7 proc. zgłoszeń, poddano moderacji. W przypadku 1312 materiałów, czyli 37,9 proc. treści, serwisy społecznościowe nie podjęły żadnych działań.

WYKRES 29. **Reakcja platform na zgłoszenia zewnętrzne**



ŹRÓDŁO: OPRACOWANIE WŁASNE

Tabela 3 przedstawia zgłoszenia według wewnętrznej kategoryzacji z podziałem na rodzaj naruszenia. Wśród wszystkich zgłoszeń zidentyfikowano:

- 1657 wpisów (47,9 proc.) zawierających zmanipulowane informacje,
- 1462 incydenty (42,3 proc.) związane z podszywaniem się pod osobę lub instytucję,
- 338 materiałów (9,8 proc.) zawierających fałszywe informacje.

TABELA 3. Rodzaj przypisanego naruszenia dla zgłoszeń zewnętrznych

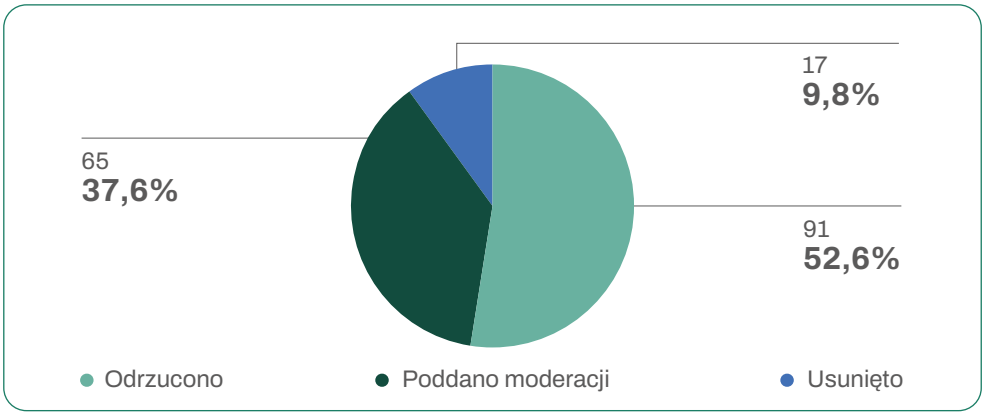
Rodzaj naruszenia	Fałszywa informacja	Manipulacja	Podszywanie się pod osobę/instytucję
Liczba zgłoszeń odrzuconych	237	805	270
Liczba zgłoszeń poddanych moderacji	12	24	56
Liczba zgłoszeń usuniętych	89	828	1 136
Całkowita liczba zgłoszeń	338	1 657	1 462

ŹRÓDŁO: OPRACOWANIE WŁASNE

X.com

Platforma X usunęła 17 wskazanych treści, czyli 9,8 proc. zgłoszeń. Bez reakcji pozostawiono 91 zaraportowanych incydentów, co odpowiada 52,6 proc. materiałów. Moderacji poddano 65 treści, czyli 37,6 proc. zgłoszeń.

WYKRES 30. Reakcja platformy X na zgłoszenia zewnętrzne

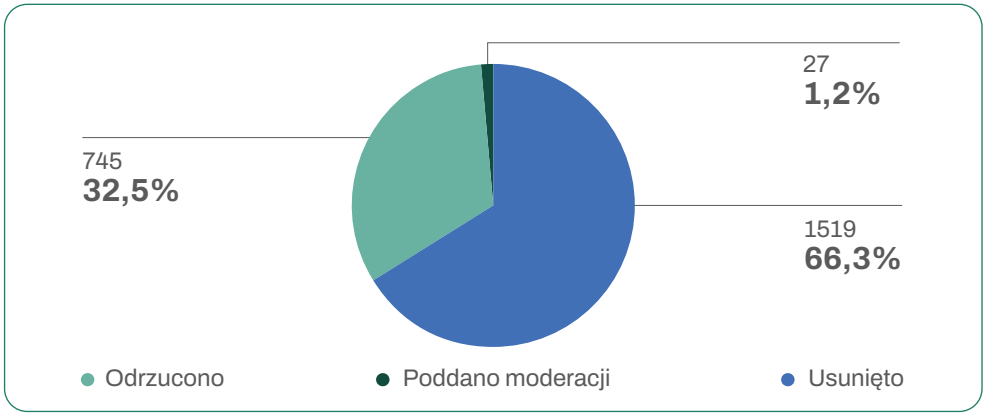


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Meta (Facebook, Instagram, Threads)

Na wykresie przedstawiono łączne dane dotyczące zaraportowanych do administracji Meta materiałów pochodzących z platform Facebook, Instagram oraz Threads. Zespoły moderujące usunęły 1519 treści, co odpowiada 66,3 proc. przypadków. Dalsze 745 zgłoszeń, czyli 32,5 proc. incydentów, zostało odrzuconych. Moderacji poddano 27 treści, a więc 1,2 proc. materiałów.

WYKRES 31. **Reakcja platformy Meta na wszystkie zgłoszenia zewnętrzne**

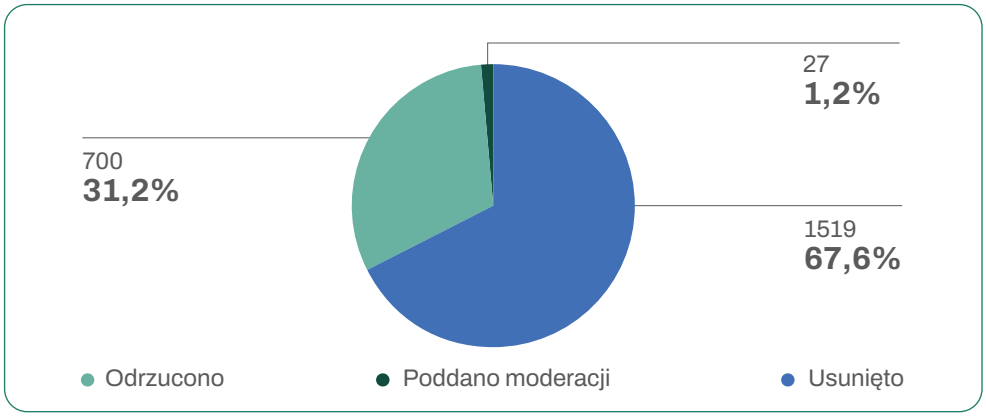


ŹRÓDŁO: OPRACOWANIE WŁASNE

Facebook

Zespół moderujący platformy Facebook usunął 1519 zgłoszonych treści, co stanowi 67,6 proc. incydentów. Odrzuconych zostało z kolei 700 zgłoszeń, co odpowiada 31,2 proc. materiałów, a 27 incydentów, czyli 1,2 proc. zgłoszonych treści, zostało poddanych moderacji.

WYKRES 32. **Reakcja platformy Facebook na zgłoszenia zewnętrzne**

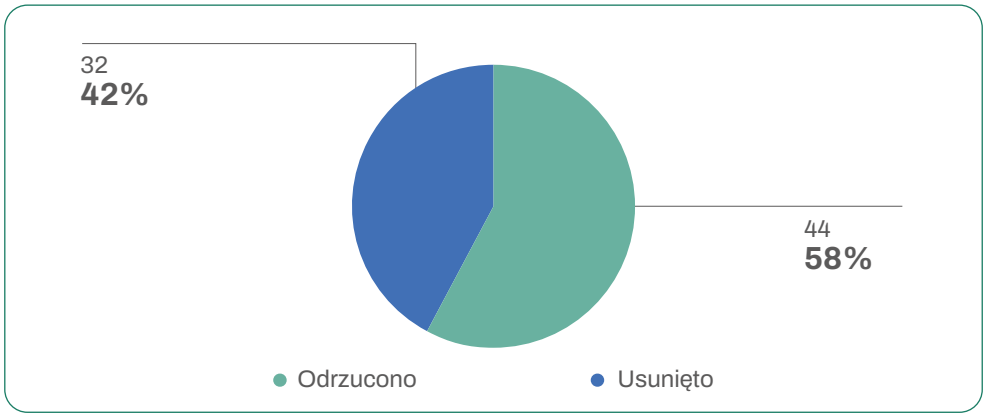


ŹRÓDŁO: OPRACOWANIE WŁASNE

Instagram

Administracja platformy Instagram usunęła 32 materiały, a więc 42 proc. incydentów. Odrzuciła 44 zgłoszenia, czyli 58 proc. treści.

WYKRES 33. **Reakcja platformy Instagram na zgłoszenia zewnętrzne**



ŹRÓDŁO: OPRACOWANIE WŁASNE

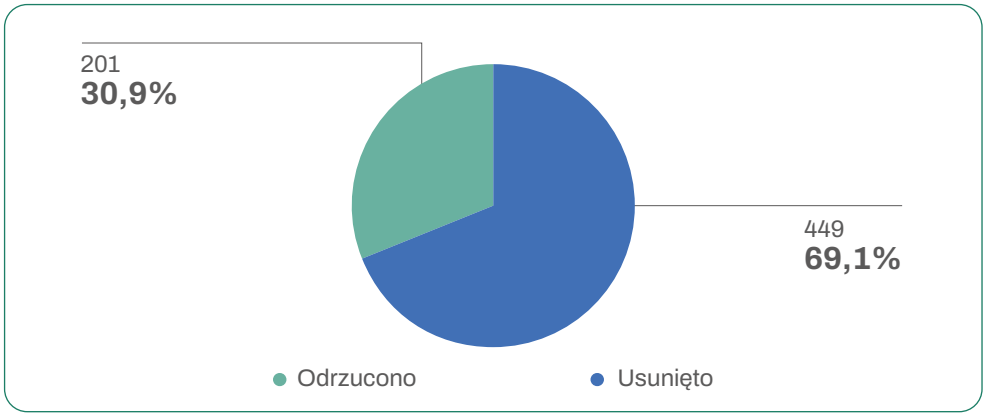
Threads

Administratorzy serwisu Threads odrzucili zgłoszenie 1 materiału.

TikTok

W raportowanym okresie platforma TikTok usunęła 449 materiałów, co przekłada się na 69,1 proc. zgłoszonych treści. Bez reakcji pozostawiono 201 treści, co stanowi 30,9 proc. zgłoszeń.

WYKRES 34. **Reakcja platformy TikTok na zgłoszenia zewnętrzne**

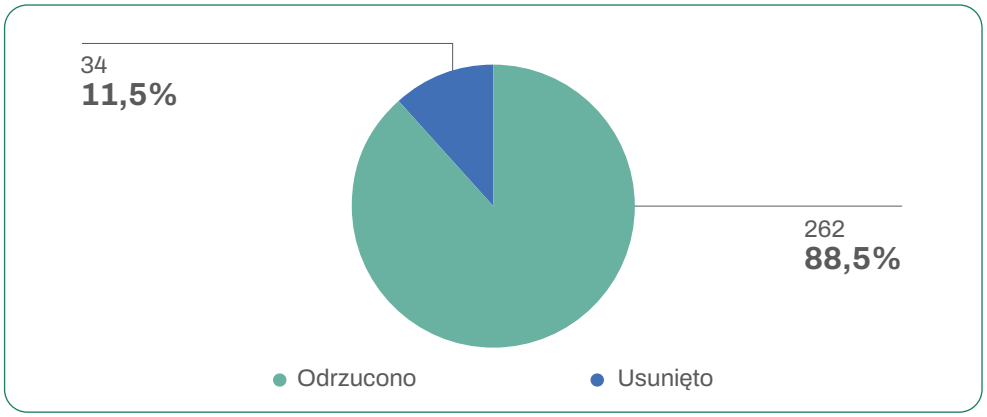


ŹRÓDŁO: OPRACOWANIE WŁASNE

Platformy Google (YouTube, Google Ads)

W przypadku zgłoszeń zewnętrznych związanych z serwisami Google, platformy te usunęły 34 treści (11,5 proc. zgłoszeń). W przypadku 262 treści (88,5 proc. zgłoszeń) zespoły moderujące nie podjęły żadnych działań.

WYKRES 35. **Reakcja platformy Google na zgłoszenia zewnętrzne**

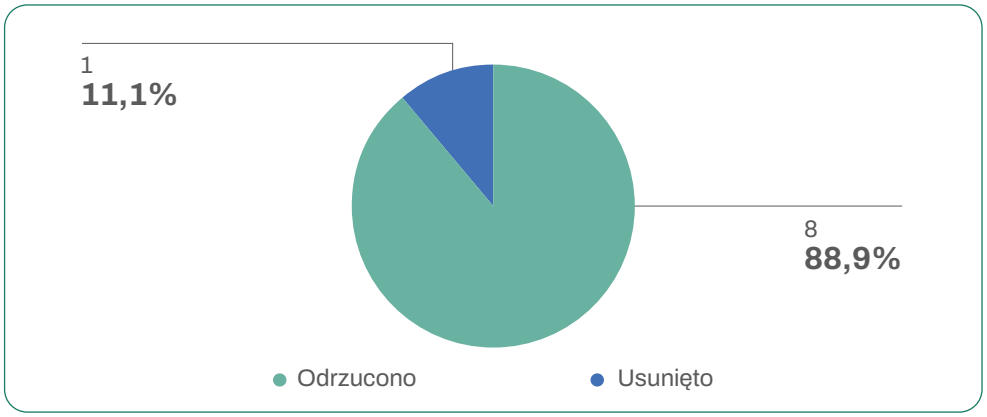


ŹRÓDŁO: OPRACOWANIE WŁASNE

Telegram

Do serwisu Telegram skierowano 9 zgłoszeń zewnętrznych. Zespół moderujący usunął 1 treść, co odpowiada 11,1 proc. zgłoszeń, natomiast bez reakcji pozostawiono 8 incydentów, czyli 88,9 proc materiałów.

WYKRES 36. **Reakcja platformy Telegram na zgłoszenia zewnętrzne**



ŹRÓDŁO: OPRACOWANIE WŁASNE

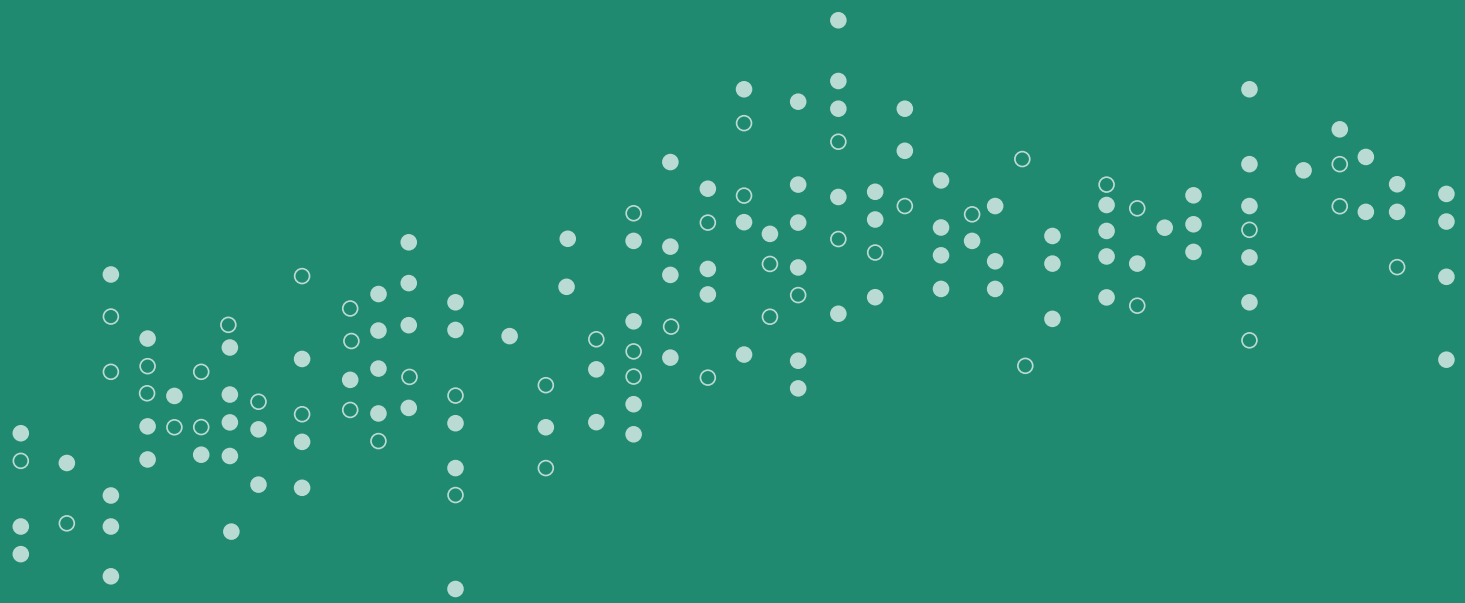
Discord

Administracja serwisu Discord odrzuciła wszystkie 4 zgłoszenia, które zostały do niej skierowane.

Inne

Do pozostałych platform zgłoszono 2 publikacje, z których 1 incydent zakończył się usunięciem treści, a 1 materiał pozostał bez reakcji.

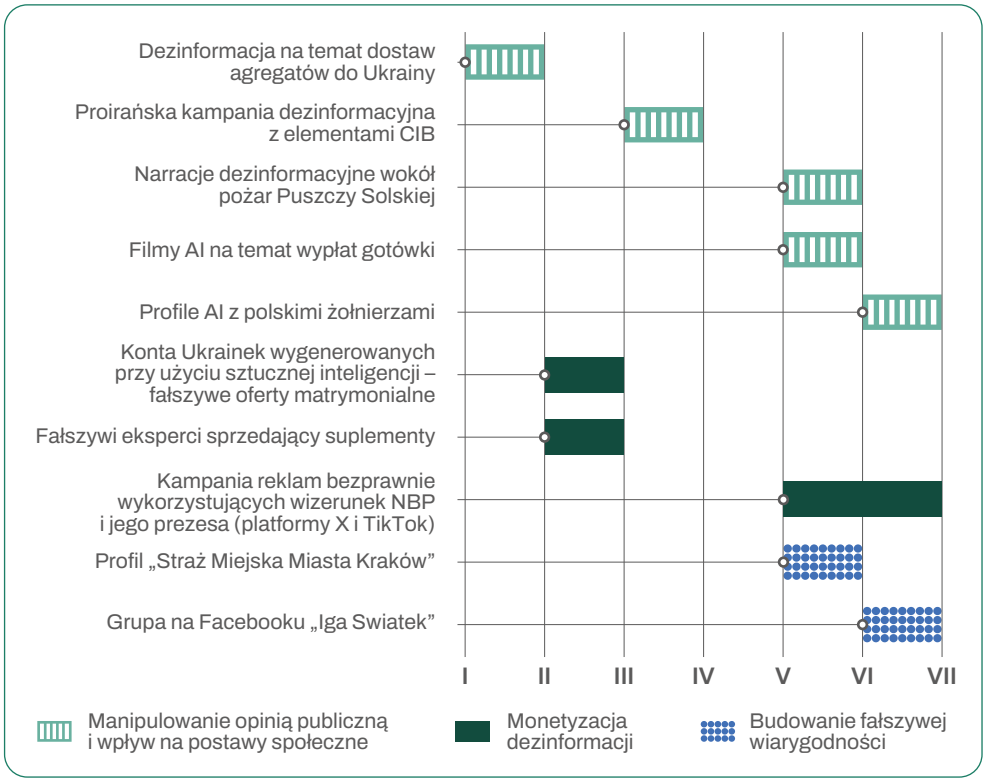
Case studies – narracje dezinformacyjne



Case studies – narracje dezinformacyjne

W tym rozdziale zostały przedstawione case studies z analizowanego okresu. Przykłady opisano, dzieląc je ze względu na prawdopodobny cel realizowanej dezinformacji. Przypadki podzielone zostały na 3 kategorie: manipulowanie opinią publiczną i wpływ na postawy społeczne (na wykresie numer 37 oznaczone zielonym paskowaniem), monetyzacja dezinformacji (kolor ciemny zielony) oraz budowanie fałszywej wiarygodności (kolor niebieski).

WYKRES 37. **Miesiąc występowania prezentowanych treści**



ŹRÓDŁO: OPRACOWANIE WŁASNE

Manipulowanie opinią publiczną i wpływ na postawy społeczne

Pierwsza kategoria obejmuje treści dezinformacyjne, których celem było wpływanie na społeczeństwo i kształtowanie jego opinii. Autorzy i autorki publikacji wykorzystywali fałszywe lub wprowadzające w błąd przekazy, aby wzmacniać określone narracje dotyczące sytuacji w kraju i na świecie.

Dezinformacja na temat dostaw agregatów do Ukrainy

W styczniu 2026 r. analitycy i analityczki OAD NASK zidentyfikowali powtarzalne przekazy odnoszące się do dostaw agregatów prądotwórczych z Polski do Ukrainy. Treści dotyczyły zarówno wsparcia realizowanego przez instytucje państwowe, jak i w ramach akcji „Ciepło z Polski dla Kijowa", zainicjowanej przez Fundację Stand with Ukraine. Zgłoszone wpisy koncentrowały się wokół dwóch głównych tez dezinformacyjnych, które funkcjonują w polskiej przestrzeni informacyjnej od początku wybuchu wojny w Ukrainie.

1. **Pomoc dla Ukrainy odbywa się kosztem Polek i Polaków**

   W ramach tej narracji rozpowszechniano treści sugerujące, że przekazywanie agregatów Ukrainie odbywa się kosztem potrzeb obywateli Polski, którym brakuje środków na zakup opału. W tym kontekście pojawiały się wpisy zestawiające rzekomą sytuację Polek i Polaków (na przykład „Polacy marzną") z pomocą kierowaną do Ukrainy. Takie treści, zawierające zmanipulowany kontekst oraz wskazujące fałszywą przyczynę niekorzystnych zjawisk, miały na celu wywołanie poczucia niesprawiedliwości i wzbudzenie negatywnych emocji zarówno wobec działań państwa, jak i wobec obywateli Ukrainy. Narracja ta była wzmacniana dodatkowymi wątkami o charakterze politycznym i tożsamościowym (na przykład odniesieniami do rzekomych sposobów, w jakie Ukraina miała być uprzywilejowana w kontaktach z Polską, czy zagrożeń dla naszego państwa ze strony Ukraińców i Ukrainek). Treści te pojawiały się zarówno w mediach społecznościowych, jak i na zewnętrznych stronach internetowych.

OBRAZ 3. Materiał zawierający manipulację na temat wysyłki agregatów do Ukrainy



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 4. Grafika wygenerowana za pomocą AI związana z przekazywaniem generatorów Ukrainie

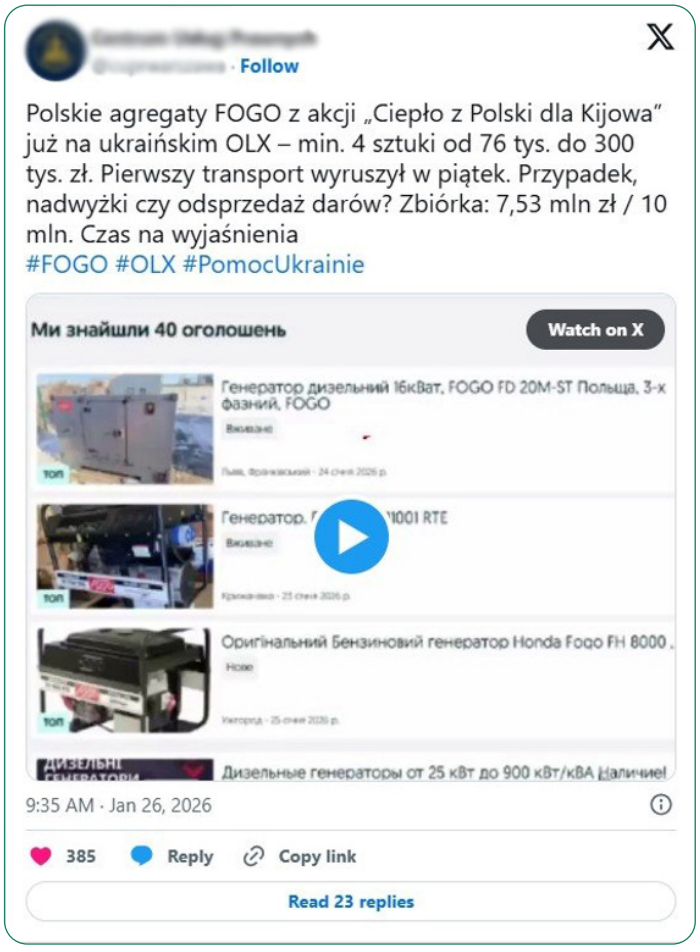


ŹRÓDŁO: FAŁSZYWY PORTAL WWW

2. Ukraińcy i Ukrainki wzbogacają się na pomocy z Polski

Druga zidentyfikowana narracja koncentrowała się na twierdzeniach, że przekazywana pomoc jest wykorzystywana przez Ukrainki i Ukraińców w celach komercyjnych. W analizowanych wpisach sugerowano, że agregaty trafiają na ukraińskie platformy aukcyjne i są odsprzedawane, co ma prowadzić do osiągania zysków przez stronę ukraińską. Tezy te opierały się głównie na niezweryfikowanych obserwacjach, takich jak samo pojawienie się ofert sprzedaży[3] oraz insynuacjach dotyczących pochodzenia sprzętu. Nie przedstawiano w nich wiarygodnych dowodów potwierdzających związek między przekazaną pomocą a ogłoszeniami z platform sprzedażowych.

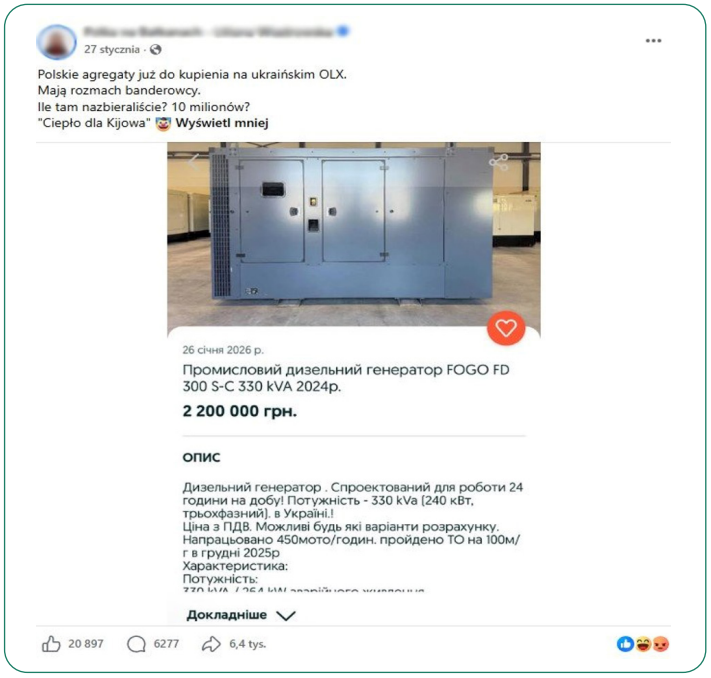
OBRAZ 5. **Wpis sugerujący, że agregaty z polskich akcji pomocowych są sprzedawane na ukraińskich aukcjach**



ŹRÓDŁO: X.COM

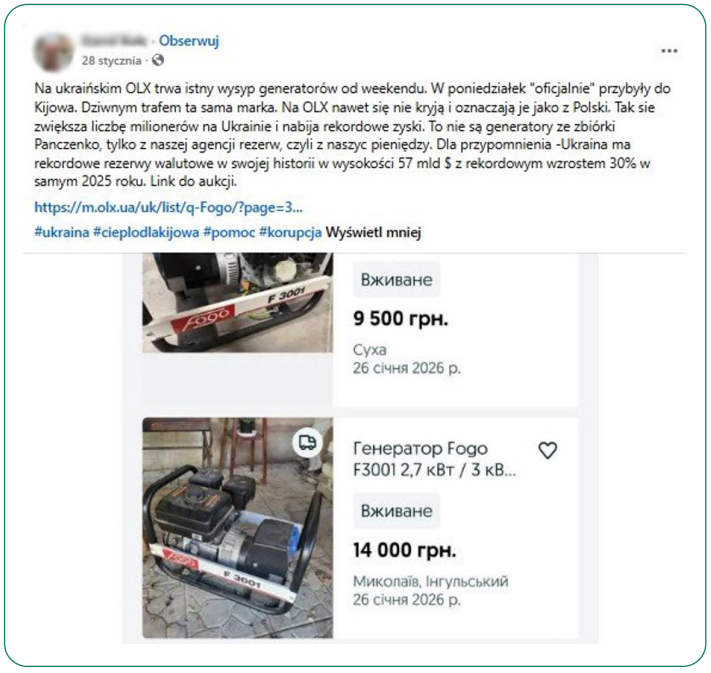
3 dezinformujący wpis na portalu X [dostęp: 17.08.2026 r., godz. 13.28]

OBRAZ 6. **Wpis sugerujący, że agregaty z polskich akcji pomocowych są sprzedawane na ukraińskich aukcjach**



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 7. **Wpis sugerujący, że agregaty z polskich akcji pomocowych są sprzedawane na ukraińskich aukcjach**



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 8. Wpis sugerujący, że agregaty z polskich akcji pomocowych są sprzedawane na ukraińskich aukcjach



ŹRÓDŁO: YOUTUBE.COM

Wszystkie zgłoszenia dotyczące dezinformacji na temat agregatów prądotwórczych zostały odrzucone, niezależnie od tego na której platformie zostały umieszczone wskazane w nich wpisy (Facebook, X, YouTube).

Proirańska kampania dezinformacyjna z elementami CIB

W marcu 2026 r. analitycy i analityczki OAD NASK zaobserwowali wzrost liczby wpisów o charakterze proirańskim, rozpowszechnianych szczególnie na platformie X i powielanych na platformie Instagram, które odnosiły się do sytuacji geopolitycznej na Bliskim Wschodzie. Treści te publikowane były przez fałszywe konta stylizowane na profile informacyjne, wojskowe lub oficjalne kanały państwowe.

Skala, powtarzalność oraz forma publikowanych materiałów wskazują na możliwy charakter skoordynowanych działań nieautentycznych (CIB), ukierunkowanych na zwiększenie zasięgu wybranych narracji oraz kształtowanie u odbiorców określonego obrazu konfliktu USA i Izraela z Iranem. Wpisy

często wykorzystywały identyczne lub bardzo zbliżone schematy komunikacyjne (na przykład oznaczenie „BREAKING", przypisywanie wypowiedzi przywódcom państw, powielanie tych samych treści w różnych wersjach językowych).

Zaobserwowano również wykorzystanie narzędzi do automatyzacji publikacji i elementów generowanych przy użyciu sztucznej inteligencji, na przykład powielanie podobnych grafik, tłumaczenie wpisów na różne języki, komunikaty przypisywane osobom publicznym w postaci wygenerowanych deepfake'ów. Zastosowanie takich rozwiązań umożliwia szybkie rozpowszechnianie przekazu, zwiększenie jego zasięgu oraz nadanie mu pozorów wiarygodności, co utrudnia odbiorcom i odbiorczyniom odróżnienie treści autentycznych od zmanipulowanych.

Platformy w zdecydowanej większości odrzuciły zgłoszenia Ośrodka dotyczące wskazanych wyżej kont i wpisów.

Niektóre narracje, które pojawiały się w tej kampanii dezinformacyjnej, zostały omówione poniżej.

1. Benjamin Netanjahu nie żyje

Jednym z dominujących wątków były fałszywe lub niezweryfikowane informacje dotyczące zamachów na premiera Izraela Benjamina Netanjahu bądź jego śmierci. W analizowanych wpisach sugerowano, że zginął w wyniku ataków lub informowano o planowanych działaniach militarnych przeciwko miejscu jego zamieszkania. Treści te miały charakter alarmistyczny i były często oznaczane jako pilne wiadomości. Miało to na celu najprawdopodobniej zwiększenie wiarygodność przekazu i skłonienie odbiorców i odbiorczyń do dalszego rozpowszechniania wprowadzających w błąd materiałów. W wielu przypadkach wykorzystywano także zmanipulowane grafiki oraz fałszywe komunikaty stylizowane na doniesienia medialne.

OBRAZ 9. Wpis o ataku na Benjamina Netanjahu

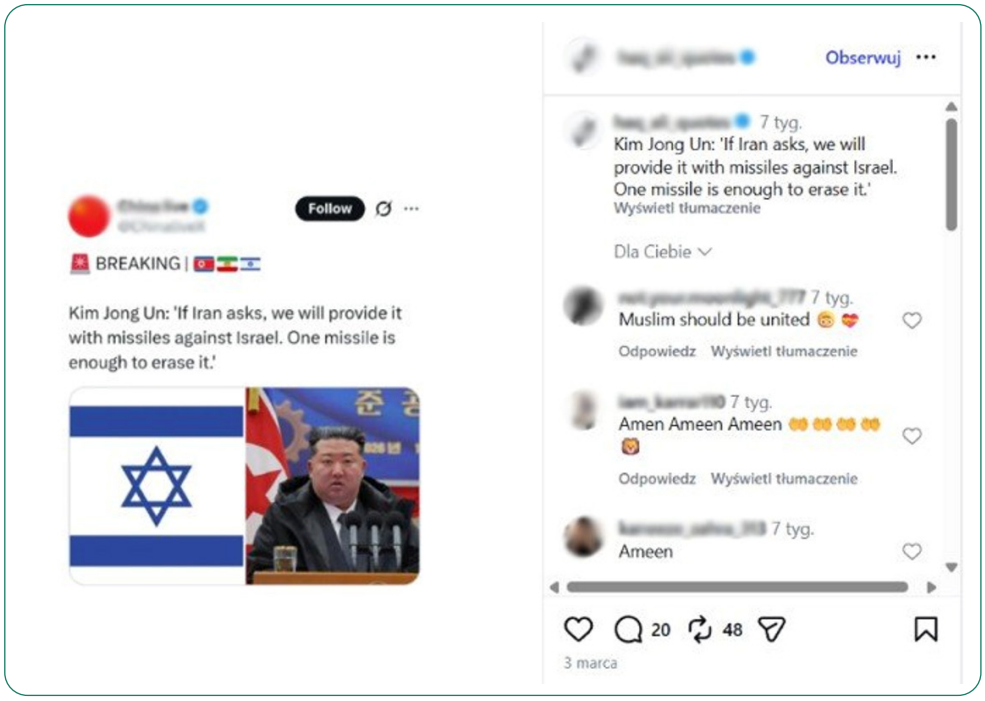


ŹRÓDŁO: INSTAGRAM.COM

2. Korea Północna i Chiny wspierają militarnie Iran

Druga narracja koncentrowała się na przedstawianiu Korei Północnej jako militarnego sojusznika Iranu w konflikcie z USA i Izraelem. W analizowanych wpisach przywódcy Korei Północnej Kim Dzong Unowi przypisywano wypowiedzi sugerujące gotowość wsparcia Iranu, w tym również użycie broni nuklearnej lub innych środków prowadzących do eskalacji konfliktu. Treści te miały charakter alarmistyczny, a ich celem było wzmacnianie poczucia zagrożenia oraz przedstawienie konfliktu jako globalnej konfrontacji. W analizowanych treściach pojawiały się również odniesienia do Chin jako potencjalnego uczestnika konfliktu lub sojusznika Iranu.

OBRAZ 10. Wpis sugerujący militarne wsparcie Iranu przez Koreę Północną



ŹRÓDŁO: INSTAGRAM.COM

OBRAZ 11. Wpis sugerujący militarne wsparcie Iranu przez Chiny



ŹRÓDŁO: INSTAGRAM.COM

3. Ali Chamenei żyje

Równolegle rozpowszechniano treści sugerujące, że przywódca Iranu Ali Chamenei żyje, mimo jednoznacznych doniesień o jego śmierci. W analizowanych wpisach pojawiały się komunikaty, według których miał on nadal sprawować władzę. Przekazy miały na celu przedstawienie Iranu jako stabilnego państwa.

OBRAZ 12. **Wpis sugerujący, że Ali Chamenei wciąż żyje**



ŹRÓDŁO: FACEBOOK.COM

Narracje dezinformacyjne wokół pożaru Puszczy Solskiej

5 maja 2026 roku doszło do rozległego pożaru znajdującej się na terenie województwa lubelskiego Puszczy Solskiej. Był to największy pożar lasu w Polsce od ponad 30 lat. 15 maja 2026 roku Ministerstwo Klimatu i Środowiska oficjalnie ogłosiło zakończenie akcji gaśniczej. Pożar objął łącznie 885 hektarów, w tym obszar rezerwatu „Długie bagno" oraz terenów wchodzących w skład sieci Natura 2000[4].

8 maja 2026 roku Prokuratura Rejonowa w Biłgoraju wszczęła śledztwo w sprawie pożaru Puszczy Solskiej. Tego samego dnia Rzecznik Prokuratury Okręgowej w Zamościu Rafał Kawalec poinformował, że nie ma żadnych obiektywnych dowodów, które wskazywałyby na umyślne działania osób trzecich, jednakże zastrzegał, że możliwość przeprowadzenia oględzin pozostawała ograniczona ze względu na trwającą akcję straży pożarnej[5].

Mimo to w odniesieniu do wydarzeń na Lubelszczyźnie w mediach społecznościowych odnotowano rozpowszechnianie licznych narracji dezinformacyjnych. Dotyczyły one między innymi rzekomego umyślnego podpalenia lasu przez obywateli Ukrainy oraz tego, jakoby pożar Puszczy Solskiej miał stanowić narzędzie celowego „czyszczenia terenu" pod bazę NATO.

1. Puszcza Solska została podpalona przez Ukraińców

Główną osią narracji była zmanipulowana interpretacja wypowiedzi ukraińskiej aktywistki Natalii Panczenko, która pojawiła się w wywiadzie dla ukraińskiej telewizji Kanał 5 w lutym 2025 roku[6]. Działaczka powiedziała wtedy (w kontekście trwającej wówczas kampanii prezydenckiej w Polsce), że pogłębianie polaryzacji i podziałów między Polakami a Ukraińcami może stanowić zagrożenie przede wszystkim dla Polski, a wskutek niego na terytorium naszego kraju miałoby dochodzić do bójek czy podpaleń.

Aktorzy dezinformacji publikujący treści łączące pożar na Lubelszczyźnie z działalnością obywateli Ukrainy często odnosili się do tej wypowiedzi. Wskazywano, że podpalenia były przez N. Panczenko „zapowiadane" i obecnie ten „plan" jest realizowany. Przekaz wzmacnia nastroje antyukraińskie w kraju, pogłębia polaryzację oraz może wywoływać poczucie zagrożenia ze strony obywateli Ukrainy.

4 https://www.gov.pl/web/klimat/bilans-pozaru-w-puszczy-solskiej--trwa-szacowanie-strat-i-planowanie-odnowy-lasu [dostęp: 17.08.2026 r., godz. 13.30]

5 https://www.pap.pl/aktualnosci/pozar-w-puszczy-solskiej-prokuratura-wszczela-sledztwo [dostęp: 19.08.2026 r., godz. 09.23]

6 https://www.youtube.com/watch?v=IuNxGKq9sng&t [dostęp: 17.08.2026 r., godz. 13.32]

Nie przedstawiano żadnych dowodów na poparcie tezy. Sformułowano ją wyłącznie na podstawie wypowiedzi ukraińskiej aktywistki. Promujące ją wpisy, pomimo zgłoszeń ze strony OAD NASK, nadal są w większości dostępne na platformach społecznościowych.

OBRAZ 13. **Wpis na temat rzekomych podpaleń**



ŹRÓDŁO: X.COM

2. Pożar Puszczy Solskiej to zamierzone działanie w celu utworzenia bazy NATO

W ramach tej narracji aktorzy dezinformacji próbowali przekonać odbiorców i odbiorczynie, że pożar Puszczy Solskiej nie był przypadkowy. Miał być zamierzoną akcją, której celem było wypalenie ziemi i przygotowanie terenu pod przyszłą infrastrukturę wojskową NATO. Spekulowano, że tego typu działania są zapowiedzią przyszłego konfliktu zbrojnego na wschodniej granicy Polski. Pojawiały się również sugestie, że rzekoma operacja jest inspirowana działaniami Izraela w Strefie Gazy, Palestynie i Libanie.

Narracja pojawiała się na najpopularniejszych platformach społecznościowych w polskojęzycznej przestrzeni informacyjnej. W jej ramach udostępniana była między innymi grafika prezentująca rzekome plany związane z bazą NATO. W prawym dolnym rogu obrazu widoczne jest logo Gemini – dużego modelu językowego firmy Google (por. obraz 14). Prawdopodobnie właśnie tego narzędzia użyto do wygenerowania lub zmanipulowania ilustracji. Pojawianie się obrazu na różnych platformach, takich jak Facebook, X, TikTok czy Bluesky, może sugerować, że był to element skoordynowanej kampanii.

Promowana teoria o ukrytych działaniach mających na celu budowę infrastruktury NATO na Lubelszczyźnie wpisuje się w rosyjski i białoruski przekaz propagandowy. W ten sposób budowany jest obraz Sojuszu Północnoatlantyckiego jako prącego do konfliktu oraz prowokującego Białoruś

i Rosję poprzez umieszczanie baz blisko granic tych państw. W skrajnym scenariuszu tego rodzaju narracje mogłyby być wykorzystywane przez Mińsk i Moskwę do usprawiedliwiania swoich potencjalnych działań przeciwko Polsce.

Platformy sprawnie reagowały na zgłoszenia dotyczące tej narracji o pożarze Puszczy Solskiej. Większość wskazanych przez OAD NASK treści nie jest już dostępna.

OBRAZ 14. **Wpis o rzekomym „czyszczeniu" terenu pod bazę NATO**

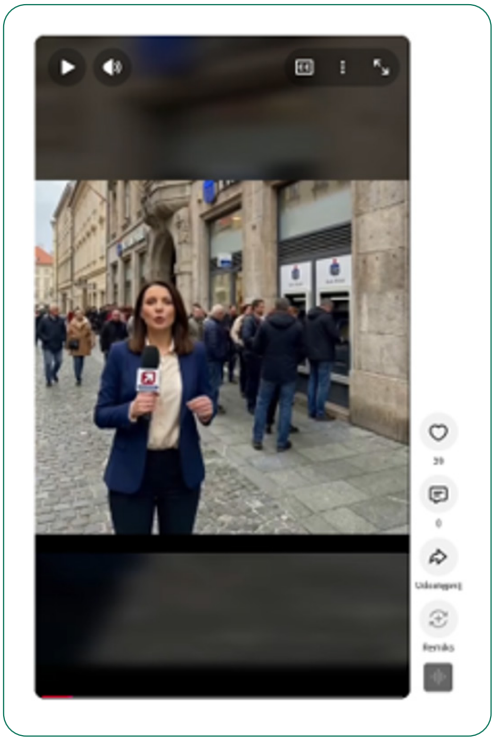


ŹRÓDŁO: TIKTOK.COM

Filmy AI na temat wypłat gotówki

W maju 2026 roku na platformie YouTube pojawiły się filmy wygenerowane za pomocą AI, w których sugerowano, że Polki i Polacy masowo wypłacają gotówkę z bankomatów. Powodem miałby być strach przed możliwym kryzysem gospodarczym. Jeden z materiałów był stylizowany na relację telewizyjną.

OBRAZ 15. **Deepfake o wypłatach z bankomatów**

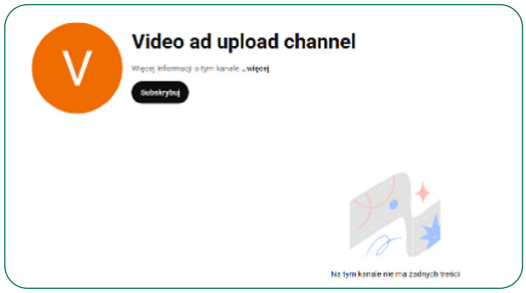


ŹRÓDŁO: YOUTUBE.COM

Zespół OAD NASK regularnie identyfikuje przekazy dezinformacyjne dotyczące rzekomych ograniczeń w wypłacaniu gotówki. Narracje te przypisują ich wprowadzenie Unii Europejskiej lub wiążą je z innymi okolicznościami, takimi jak kryzys gospodarczy czy zaangażowanie Polski w wojnę. Pierwszy raz zaobserwowano jednak narrację tego rodzaju w formie materiału wygenerowanego przy pomocy sztucznej inteligencji. Choć jakość filmu i poziom jego realizmu były niskie, tego typu treści mogą wywołać w przyszłości panikę i zmniejszenie zaufania do instytucji finansowych.

W ramach analizy zgłoszonych materiałów zaobserwowano, że deepfake o wypłatach gotówki nie jest już dostępny na platformie YouTube, w przeciwieństwie do kanału, na którym go opublikowano. Kanał nie zawiera obecnie żadnych treści.

OBRAZ 16. Kanał, który opublikował materiał AI o wypłatach z bankomatów

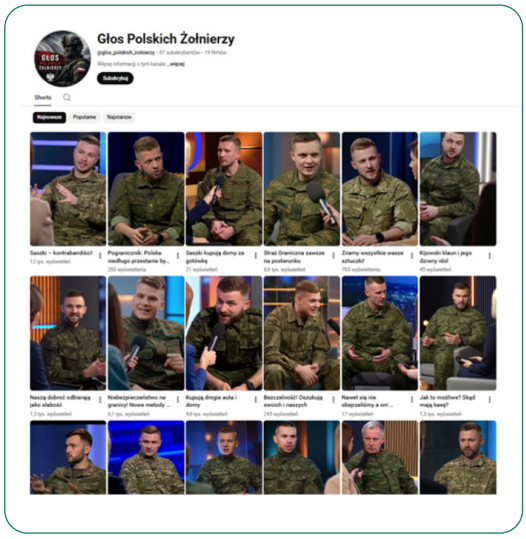


ŹRÓDŁO: YOUTUBE.COM

Profile AI z polskimi żołnierzami

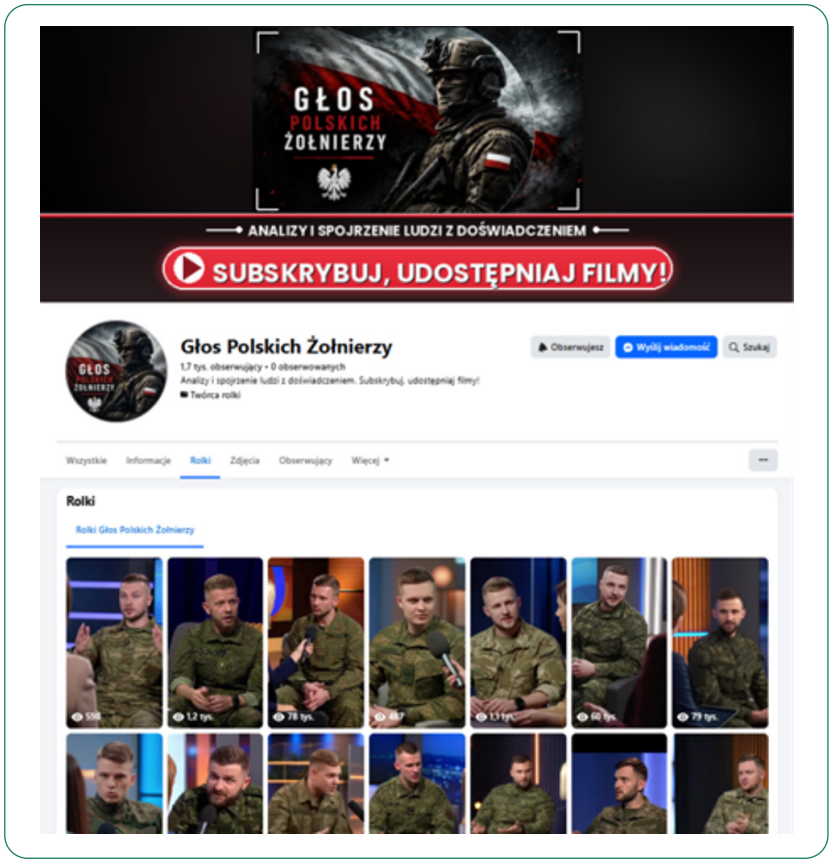
6 czerwca 2026 roku do OAD NASK przesłano informację o koncie Głos Polskich Żołnierzy w serwisie YouTube. Tego samego dnia zostało ono zgłoszone do platformy Google. Ze względu na brak reakcji zgłoszenie ponawiano jeszcze dwukrotnie. Platforma usunęła zgłaszane treści dopiero po trzecim zgłoszeniu z dnia 10 lipca 2026 roku. W wyniku dalszej analizy Ośrodek zidentyfikował kolejne konta o zbliżonym lub identycznym sposobie działania na platformach YouTube, Facebook oraz TikTok. Kanały szerzyły przekazy dezinformacyjne za pomocą wygenerowanych przy użyciu AI wypowiedzi fałszywych polskich żołnierzy. Prezentowane narracje miały charakter antyukraiński, były także wymierzone w Unię Europejską. Konta zawierały podobną treść, ale w niektórych przypadkach miały inne nazwy. Może to wskazywać na ryzyko kontynuacji lub pojawienia się nowych odsłon tej kampanii dezinformacyjnej w przyszłości.

OBRAZ 17. Kanał „Głos Polskich Żołnierzy”



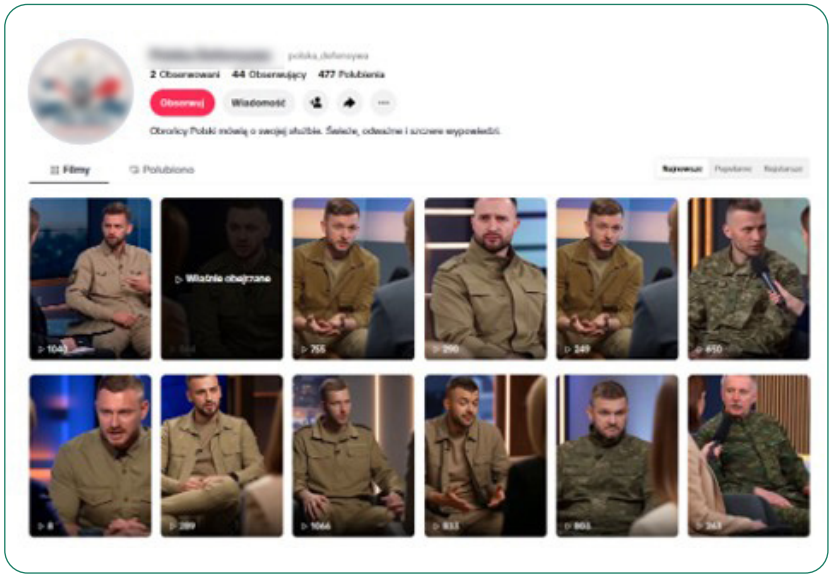
ŹRÓDŁO: YOUTUBE.COM

OBRAZ 18. Kanał „Głos Polskich Żołnierzy”



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 19. Profil AI z polskimi żołnierzami



ŹRÓDŁO: TIKTOK.COM

Materiały miały charakterystyczną formę kilkusekundowych wypowiedzi stylizowanych na wywiad telewizyjny. Większość z nich prezentowała negatywny, stereotypowy obraz Ukraińców i Ukrainek oraz Ukrainy. Ukraińscy uchodźcy i uchodźczynie byli w nich przedstawiani jako wewnętrzne zagrożenie dla Polek i Polaków, którzy stopniowo mieliby stać się obywatelami drugiej kategorii we własnym państwie. Ukraina była natomiast przedstawiana jako kraj stanowiący zagrożenie dla Polski pod względem gospodarczym i militarnym. W jednym z materiałów stwierdzono, że Polska znajduje się w niekorzystnym położeniu geopolitycznym pomiędzy Ukrainą a Niemcami, czyli dwoma państwami, które szczególne jej zagrażają. Narracje te wpisują się w szerszy trend antyukraiński, obserwowany od początku wybuchu wojny w Ukrainie w lutym 2022 roku.

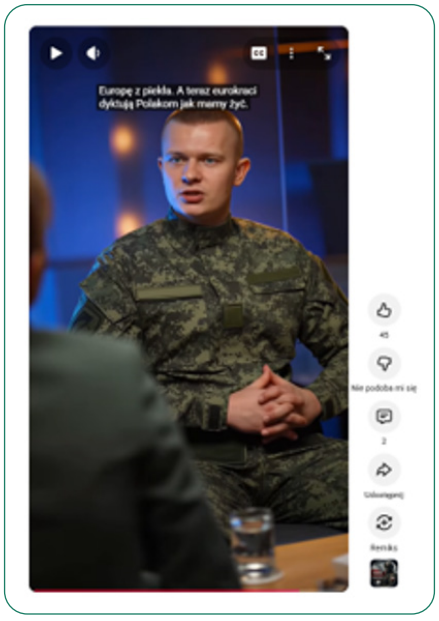
OBRAZ 20. **Deepfake z antyukraińską narracją**



ŹRÓDŁO: FACEBOOK.COM

Drugi kierunek przekazów stanowiły narracje antyunijne. Unia Europejska była w nich przedstawiana jako zagrożenie dla polskiej suwerenności i stabilności. Zamiarem Unii miałoby być osiedlenie w Polsce jak największej liczby migrantów i migrantek, by ją wynarodowić. Przekazy te wpisują się w szerszy trend antyimigranckich narracji.

OBRAZ 21. Deepfake z antyunijną narracją



ŹRÓDŁO: YOUTUBE.COM

Monetyzacja dezinformacji

Kolejna kategoria obejmuje treści, których rozpowszechnianie miało na celu uzyskanie korzyści materialnych. Autorki i autorzy wykorzystywali fałszywe profile, podszywanie się pod inne osoby, instytucje lub grupy społeczne oraz materiały tworzone przy użyciu sztucznej inteligencji.

Konta Ukrainek wygenerowanych przy użyciu sztucznej inteligencji – fałszywe oferty matrymonialne

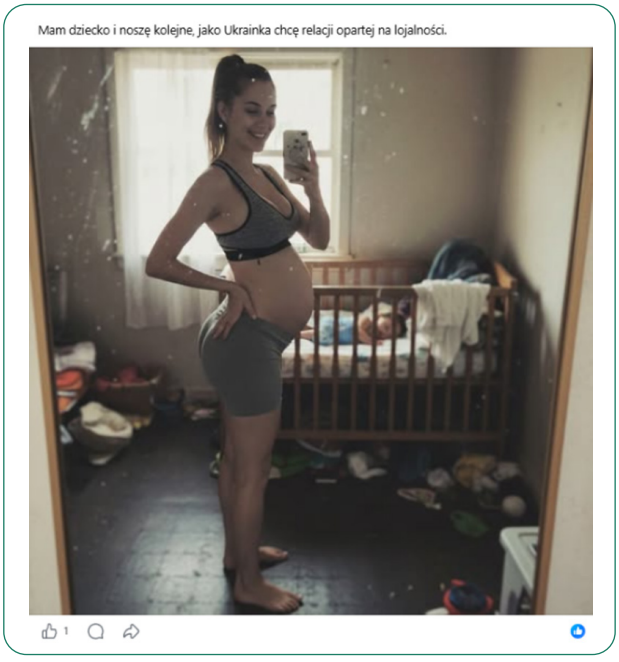
W lutym 2026 roku analityczki i analitycy OAD NASK przyjęli zgłoszenia dotyczące kont publikujących wygenerowane przy użyciu sztucznej inteligencji wizerunki kobiet w ciąży, podających się za Ukrainki poszukujące polskich partnerów. Publikacje te miały charakter manipulacyjny i wpisywały się w trend narracji mających na celu przedstawić w negatywnym świetle wizerunek Ukrainy i jej obywatelek. Skupiały się na stereotypach, sugerując, że ich obecność w Polsce motywowana jest wyłącznie chęcią nawiązywania relacji z polskimi mężczyznami.

Skala rozpowszechnienia oraz powtarzalność tych treści wskazywały na możliwy charakter skoordynowanych działań nieautentycznych (CIB), ukierunkowanych na wzmacnianie napięć społecznych oraz negatywnych postaw wobec określonej grupy narodowościowej. Wykorzystanie technologii AI

dodatkowo zwiększało wiarygodność przekazu i utrudniało jego jednoznaczną identyfikację jako manipulacji.

Działania administracji platform Meta pozostawały niespójne. Moderacja obejmowała jedynie część zgłaszanych materiałów, bez jasnego uzasadnienia kryteriów podejmowanych decyzji. Brak przejrzystości w tym zakresie skutkował pozostawieniem części treści bez reakcji, co w efekcie umożliwiało ich dalsze rozpowszechnianie oraz utrwalanie szkodliwej narracji w przestrzeni informacyjnej.

OBRAZ 22. **Materiał powstały prawdopodobnie w ramach CIB, przedstawiający wizerunek Ukrainki w ciąży, rzekomo poszukującej polskiego partnera**



ŹRÓDŁO: FACEBOOK.COM

Fałszywi eksperci sprzedający suplementy

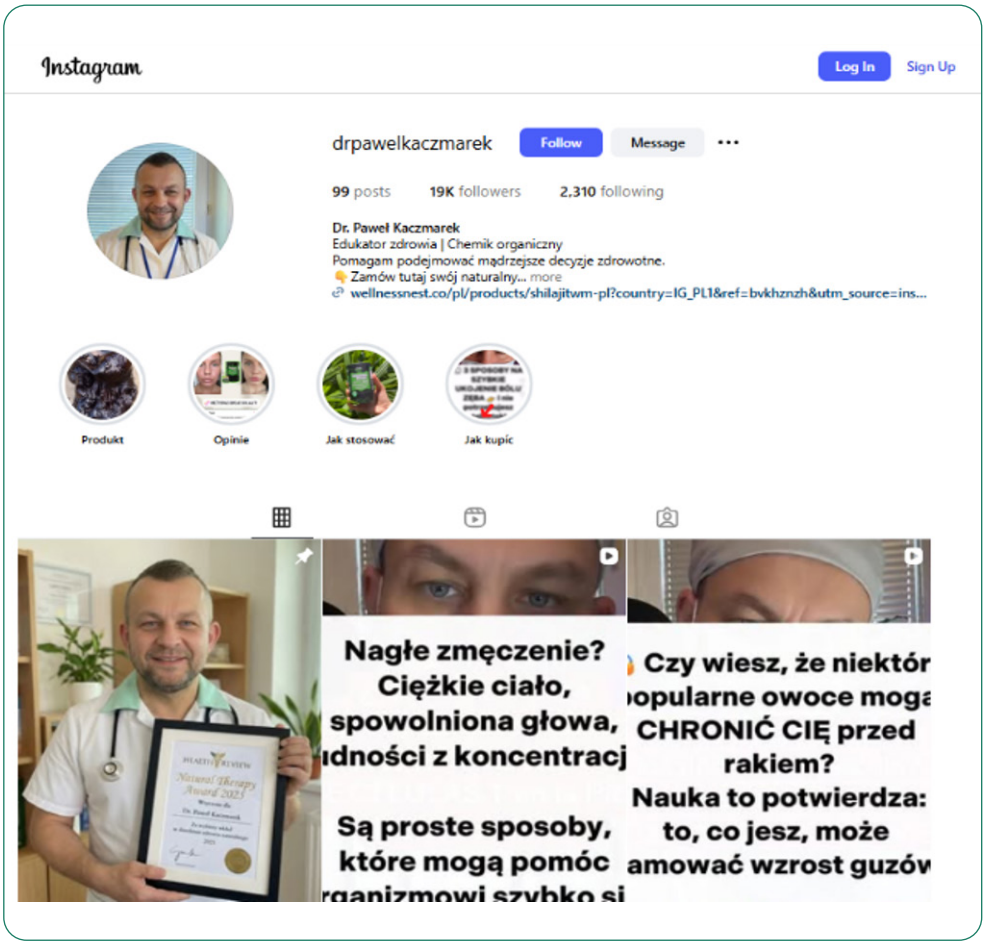
W lutym 2026 roku analityczki i analitycy OAD NASK odnotowali na platformie Instagram wzrost liczby kont podszywających się pod ekspertki i ekspertów medycznych. Profile te publikowały treści zawierające dezinformację z zakresu żywienia oraz różnych terapii, mimo że prowadzące je osoby nie posiadały odpowiednich kwalifikacji ani wykształcenia lekarskiego.

W celu zwiększenia wiarygodności przekazu wykorzystywano wizerunki osób stylizowanych na personel medyczny (na przykład za pomocą strojów lekarskich) oraz tytuł naukowy „dr", co wprowadzało użytkowników i użytkowniczki w błąd i sprzyjało budowaniu zaufania. Działania te stanowiły poważne

zagrożenie dla rzetelnej komunikacji zagadnień zdrowotnych, mogąc w konsekwencji negatywnie wpływać na zdrowie, a nawet życie odbiorców i odbiorczyń.

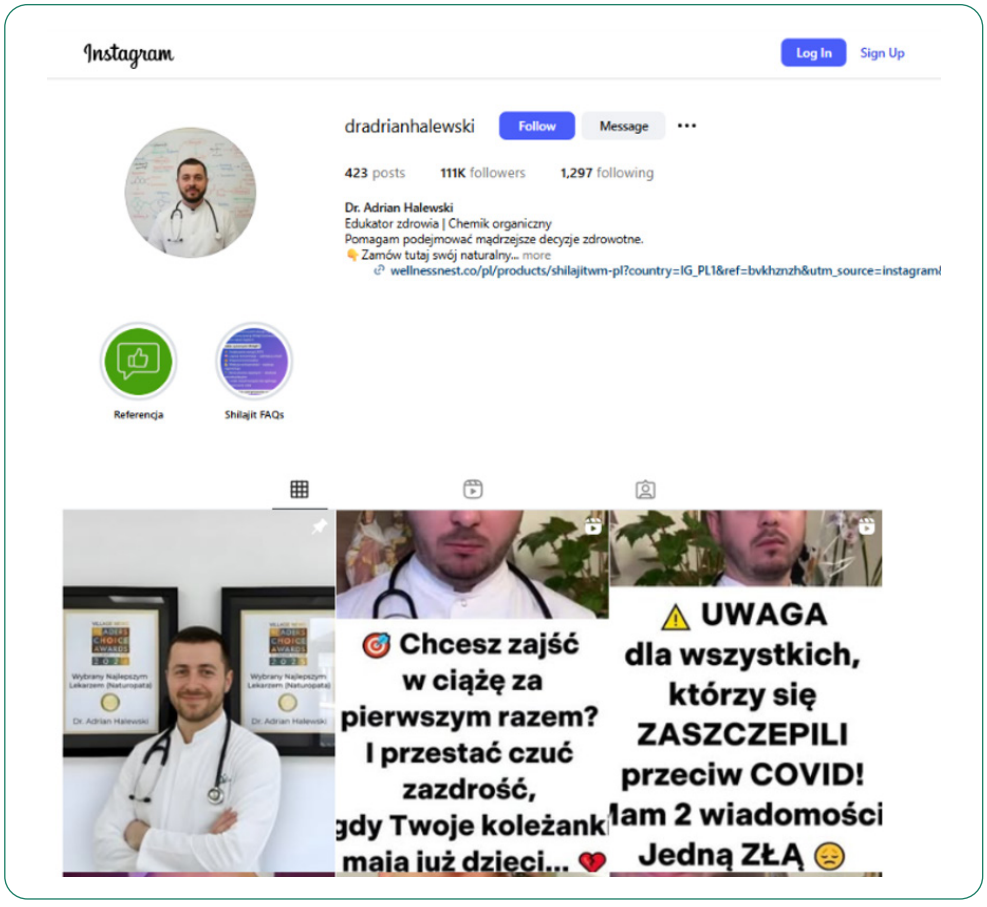
W tym przypadku platforma Meta usunęła zgłoszone konta.

OBRAZ 23. **Konto na platformie Instagram podszywające się pod eksperta medycznego**



ŹRÓDŁO: INSTAGRAM.COM

OBRAZ 24. Konto na platformie Instagram podszywające się pod eksperta medycznego



ŹRÓDŁO: INSTAGRAM.COM

Kampanie reklamowe na platformach X i TikTok

W pierwszej połowie 2026 roku na platformach X oraz TikTok zidentyfikowano kampanie reklamowe mające na celu oszustwa. Wykorzystywały one wizerunek osób powszechnie znanych, między innymi prezesa NBP Adama Glapińskiego oraz dziennikarki Moniki Olejnik. Zarówno w przypadku platformy X, jak i TikTok reklamy miały sensacyjną, clickbaitową formę. Do uwiarygodnienia oszustwa posłużył materiał stylizowany na artykuł dziennikarski, który zamieszczono na stronie imitującej portal Gazeta.pl. Dotyczył on rzekomego skandalicznego wywiadu, który Monika Olejnik miała przeprowadzić z Adamem Glapińskim.

Po kliknięciu w reklamę użytkowniczki i użytkownicy byli przekierowywani do fałszywych stron imitujących – w każdym ze zidentyfikowanych przypadków – ten sam wiarygodny portal. Prezentowane artykuły zawierały między innymi namowy do skorzystania z innowacyjnych platform inwestycyjnych, które w rzeczywistości służyły do oszustw finansowych oraz wyłudzania danych.

OBRAZ 25. Reklama zawierająca treść oszukańczą oraz bezprawne wykorzystanie wizerunku



ŹRÓDŁO: X.COM

OBRAZ 26. Reklama zawierająca treść oszukańczą oraz bezprawne wykorzystanie wizerunku



ŹRÓDŁO: TIKTOK.COM

Case studies – narracje dezinformacyjne

OBRAZ 27. Strona podszywająca się pod portal Gazeta.pl



ŹRÓDŁO: FAŁSZYWY PORTAL WWW

Podobny schemat działania na obu platformach może sugerować, że jest to skoordynowana akcja mająca na celu masowe oszustwa. Wcześniej podobne materiały pojawiały się między innymi na platformie Facebook. Większość zgłoszonych treści na platformach X, TikTok oraz Facebook została usunięta, jednak 4 materiały pozostają nadal dostępne.

Budowanie fałszywej wiarygodności

Ostatnia kategoria obejmuje treści, których celem było kreowanie fałszywej wiarygodności. W analizowanych przypadkach autorzy i autorki wykorzystali wizerunki instytucji publicznej oraz osób powszechnie znanych.

Profil „Straż Miejska Miasta Krakow”

Profil „Straż Miejska Miasta Krakow”, za pomocą którego podszywano się pod oficjalne konto służb, 12 maja 2026 roku opublikował na platformie Facebook post odnoszący się do obchodów ukraińskiego Dnia Wyszywanki . Wydarzenie odbyło się 21 maja 2026 roku na Placu Szczepańskim w Krakowie. Na udostępnionej grafice widoczne były wskazówki, dotyczące tego, jak zachowywać się w czasie trwania zgromadzenia. Wśród porad znajdowały się między innymi fałszywe zalecenia, by unikać przebywania w centrum miasta, ograniczać kontakt z uczestnikami wydarzenia oraz nie zwracać obywatelom Ukrainy uwagi w związku z zaśmiecaniem przestrzeni publicznej.

OBRAZ 28. **Wpis z poradami, jak zachowywać się podczas Dnia Wyszywanki w Krakowie**



ŹRÓDŁO: FACEBOOK.COM

Grafika z fałszywego profilu w serwisie Facebook została przeniesiona również na inne platformy społecznościowe, takie jak X. Była udostępniana między innymi przez profile znane z promowania treści antyukraińskich i antypaństwowych.

OBRAZ 29. **Wpis z fałszywymi poradami, jak zachowywać się podczas Dnia Wyszywanki w Krakowie**



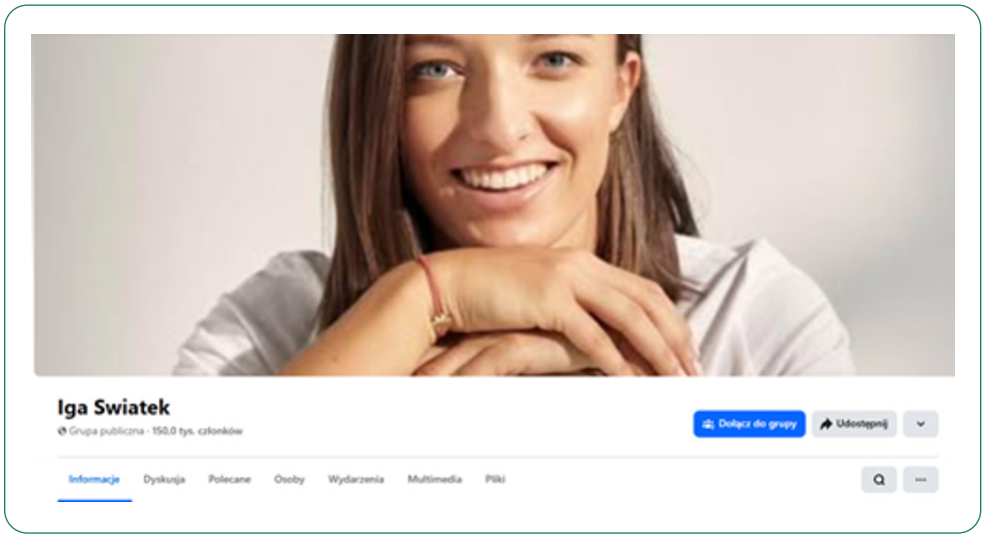
ŹRÓDŁO: X.COM

W rzeczywistości zarówno profil, jak i opublikowana na nim grafika nie były w żaden sposób autoryzowane przez krakowskie służby. Prawdziwe informacje były ogłaszane za pośrednictwem oficjalnego profilu w serwisie Facebook. Działalność fałszywego konta podważała wiarygodność Straży Miejskiej oraz kreowała negatywny wizerunek obywateli Ukrainy. Z jednej strony sugerowano, że Ukrainki i Ukraińcy przebywający w Polsce są hałaśliwi, nie dbają o porządek i lepiej unikać interakcji z nimi. Z drugiej próbowano stworzyć obraz krakowskiej Straży Miejskiej jako uległej i pozwalającej na szkodliwe zachowania. Materiały tego rodzaju mogą prowadzić do polaryzacji społecznej, spadku zaufania do służb i instytucji publicznych oraz powodować wzrost nastrojów antyukraińskich w Polsce. Profil „Straż Miejska Miasta Krakow" został usunięty przez administrację serwisu. Meta poinformowała OAD NASK o tej decyzji 18 maja, niespełna 24 godziny po wysłaniu zgłoszenia.

Grupa na Facebooku „Iga Swiatek"

W ramach monitoringu przestrzeni cyfrowej zespół OAD NASK często identyfikuje treści dezinformujące mające nakłonić odbiorców i odbiorczynie do wejścia na niebezpieczną stronę zewnętrzną poprzez załączony link. Jedną z przestrzeni do publikacji postów o takim charakterze jest publiczna grupa na Facebooku o nazwie „Iga Swiatek". Jako zdjęcia w tle grupa używa tej samej grafiki co prawdziwy profil Igi Świątek, polskiej tenisistki . Może to wywoływać u odbiorców i odbiorczyń fałszywe przeświadczenie, że jest to oficjalny kanał zrzeszający jej fanów i fanki. Ponadto administratorzy grupy mogą w ten sposób próbować dodatkowo uwierzytelnić publikowane materiały, co utrudni potencjalne podjęcie działań moderacyjnych ze strony platformy.

OBRAZ 30. **Grupa o nazwie „Iga Swiatek"**

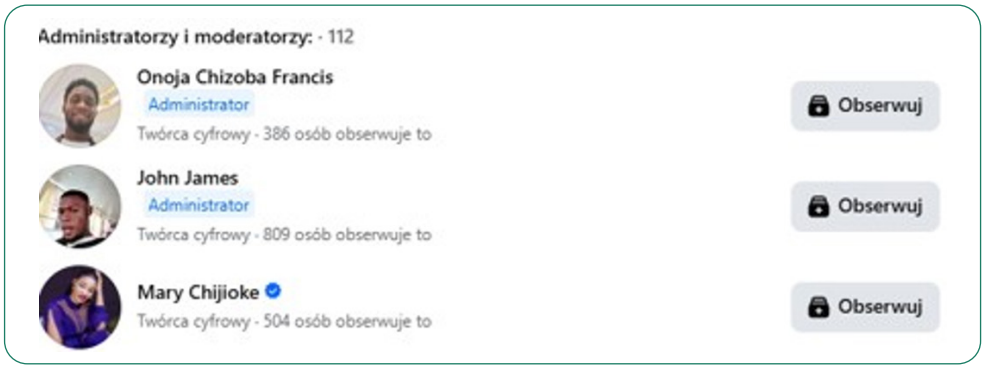


ŹRÓDŁO: FACEBOOK.COM

Grupa ma charakter otwarty. Zarządza nią 105 administratorów i administratorek oraz 7 moderatorów i moderatorek. Wszystkie te osoby noszą zagraniczne imiona i nazwiska, a lokalizacje ich kont wskazują na zamieszkanie poza Polską, na przykład w Lagos w Nigerii. Może to sugerować, że w rzeczywistości są to boty lub innego rodzaju fałszywe profile.

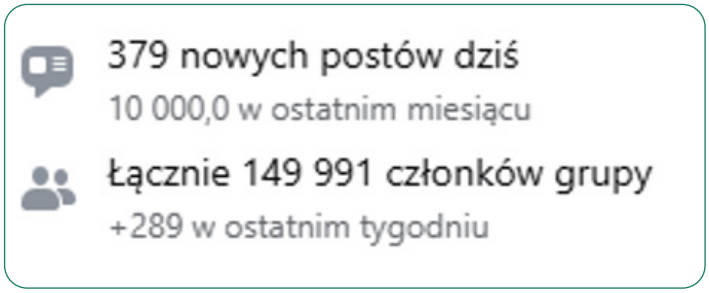
Grupa rozpoczęła swoją działalność 16 czerwca 2022 roku od publikacji wiadomości sportowych, jednak obecnie takie treści nie są już dodawane. Grupa gromadzi niemal 150 tysięcy członków i członkiń i ukazuje się w nim około 10 tysięcy postów miesięcznie. Posty na grupie są publikowane wyłącznie przez administratorów i administratorki.

OBRAZ 31. **Administratorzy grupy „Iga Swiatek”**



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 32. **Statystyki aktywności grupy „Iga Swiatek”**



ŹRÓDŁO: FACEBOOK.COM

Dodawane wpisy są typowymi clickbaitami, których celem jest wywołanie emocjonalnej reakcji odbiorców i odbiorczyń, niezależnie od prawdziwości treści kryjącej się za nagłówkiem. Komentarze wskazują przy tym na niskie zaangażowanie społeczności skupionej wokół grupy. Publikacje koncentrują się przede wszystkim na bieżących wydarzeniach oraz osobach, które w danym momencie budzą emocje i zainteresowanie.

Jednocześnie nie da się obecnie wskazać jednoznacznego celu szerzenia fałszywych treści. W zależności od aktualnych wydarzeń publikowane są głównie przekazy dotyczące osób reprezentujących środowisko polityczne. Wśród tych, których wizerunek został wykorzystany do tworzenia dezinformacji, pojawiali się między innymi Waldemar Pawlak, para prezydencka czy Jarosław Kaczyński.

OBRAZ 33. **Sensacyjny wpis w grupie „Iga Swiatek"**



ŹRÓDŁO: FACEBOOK.COM

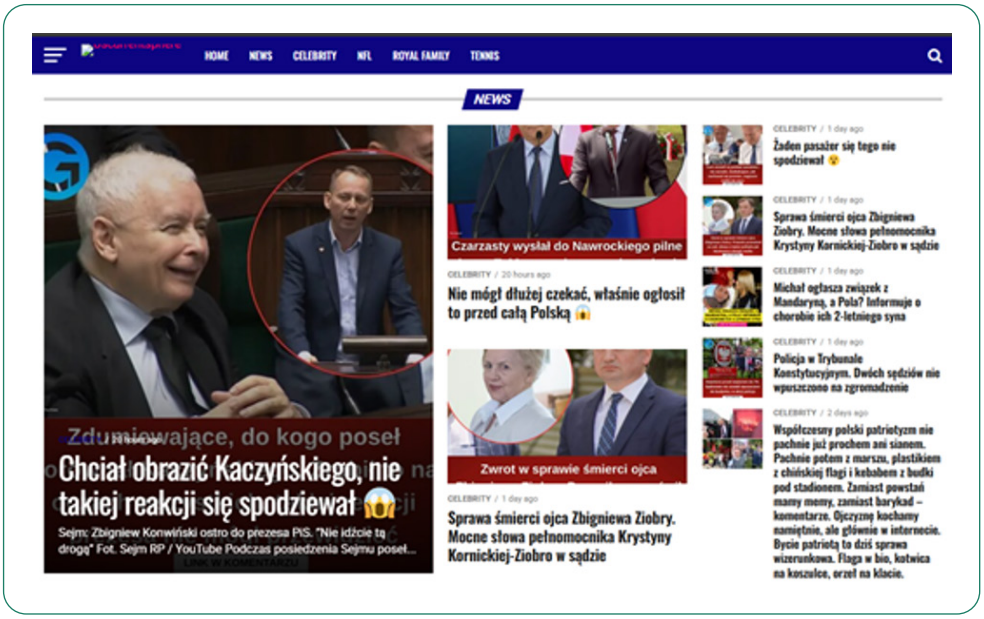
Wiele postów zawiera linki do stron podszywających się pod portale informacyjne, gdzie publikowane są fałszywe informacje. Witryny te często szybko stają się nieaktywne, jednak wkrótce zastępują je nowe, dotychczas nieużywane.

OBRAZ 34. **Sensacyjny wpis w grupie „Iga Swiatek"**



ŹRÓDŁO: FACEBOOK.COM

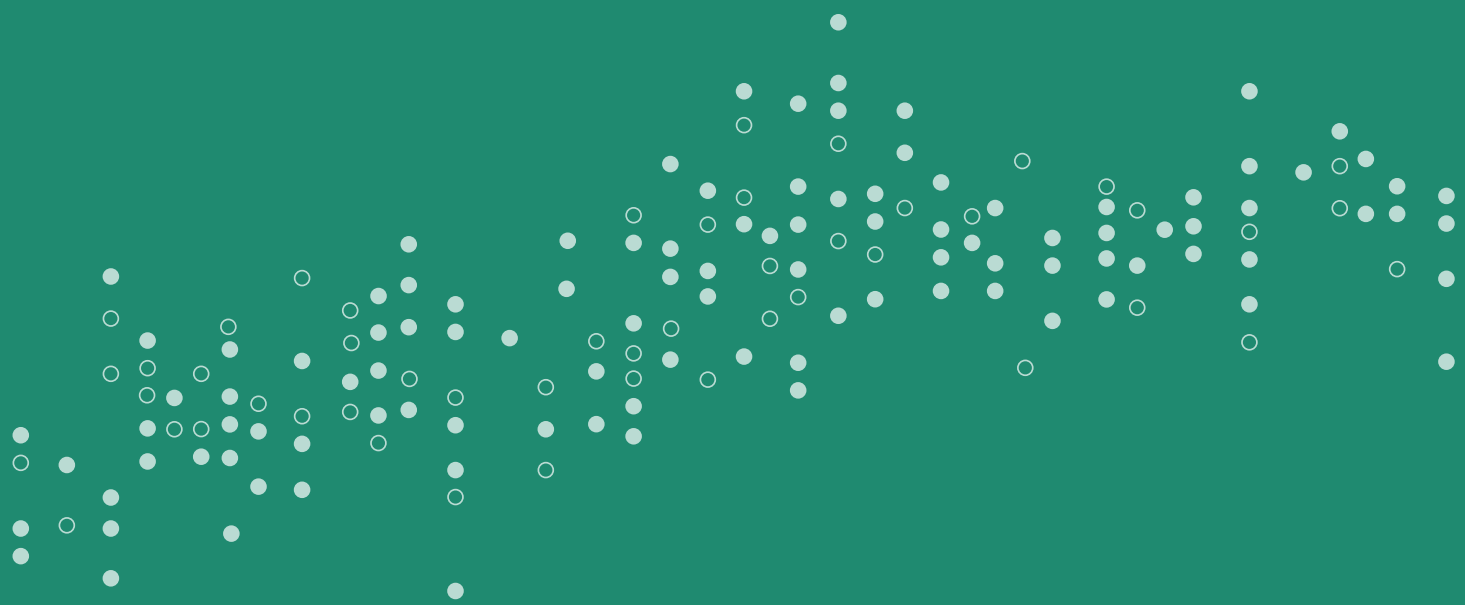
OBRAZ 35. Fałszywe portale informacyjne powiązane z grupą „Iga Swiatek”



ŹRÓDŁO: FAŁSZYWY PORTAL WWW

Platforma Facebook nie zareagowała na zgłoszenia dotyczące całej grupy, mimo że osoby, które na nią natrafiły, przesyłały je również z poziomu użytkownika przed powiadomieniem OAD NASK. Również działania Ośrodka, wraz ze szczegółowym wyjaśnieniem przypadku, nie skłoniły zespołów moderujących do działania.

Podsumowanie: współpraca z platformami – trudności i wyzwania



Podsumowanie: współpraca z platformami – trudności i wyzwania

X.com

Proces realizacji zgłoszeń przekazywanych administracji platformy X jest związany z szeregiem problemów i wyzwań.

A. Punkt kontaktowy nie zajmuje się bezpośrednią moderacją treści, a jedynie przekazuje zgłoszone materiały do zespołu odpowiedzialnego za moderację, dlatego też pracownicy i pracownice OAD NASK nie mają wpływu na to, które incydenty zostaną ostatecznie rozpatrzone. Z tego samego powodu osoby wyznaczone przez platformę X do kontaktu z instytucjami nie mają dostępu do informacji o aktualnym statusie zgłoszeń, co utrudnia analitykom i analityczkom OAD NASK monitorowanie przebiegu ich obsługi.

Bezpośrednio do zespołu moderacji trafiają wyłącznie zgłoszenia przesłane za pośrednictwem formularza Legal request submissions. Narzędzie to nie umożliwia jednak podglądu złożonych zgłoszeń ani sprawdzenia ich bieżącego statusu. Zespół OAD NASK otrzymuje jedynie informację zwrotną o tym, czy zgłoszenie zostało odrzucone, czy wskazany wpis został poddany moderacji. W treści wiadomości przesłanej na skrzynkę mailową osoby dokonującej zgłoszenia nie pojawia się uzasadnienie decyzji podjętej przez zespół moderujący, co znacząco utrudnia proces odwoławczy. Taki system zgłaszania treści jest mało efektywny i znacząco wydłuża czas obsługi.

B. Za problematyczną uznano praktykę platformy X polegającą na informowaniu autora lub autorki publikacji o złożeniu przez OAD NASK wniosku o moderację treści w momencie przekazania zgłoszenia za pośrednictwem formularza Legal request submissions. Jednostką decydującą o podjęciu działań w stosunku do przesłanych treści jest zespół moderujący samego serwisu, dlatego opisywane działanie nie ma jasnego uzasadnienia.

C. Niska skuteczność platformy X w usuwaniu treści dezinformacyjnych może wynikać z przypuszczalnego niedoboru moderatorów i moderatorek biegle posługujących się językiem polskim. Osoby znające lokalny kontekst prawdopodobnie byłyby w stanie trafniej identyfikować i skuteczniej reagować na pojawiające się zmanipulowane lub fałszywe informacje.

D. Utrzymującym się problemem pozostaje podejście platformy X do aktywności kont-botów zaangażowanych w prowadzenie zorganizowanych kampanii dezinformacyjnych. Konta te są czasowo zawieszane, a nie w pełni usuwane. Oznacza to, że w przyszłości mogą ponownie zostać wykorzystane do prowadzenia działań dezinformacyjnych.

E. Kluczowym problemem wydaje się podejście platformy X do dezinformacji, która, pomimo że jest przez nią jednoznacznie uznawana za naruszenie regulaminu, nie zawsze podlega moderacji. W konsekwencji w sytuacjach kryzysowych, wymagających szybkiej reakcji, administracja serwisu często nie podejmuje działań lub reaguje ze znacznym opóźnieniem.

Platformy Meta (Facebook, Instagram, Threads)

Proces realizacji zgłoszeń przekazywanych administracji platform Meta jest związany z szeregiem problemów i wyzwań.

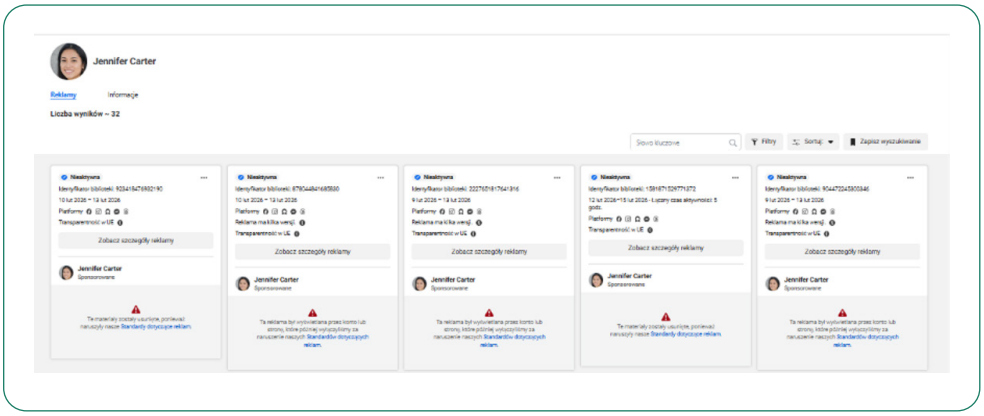
A. Zdarza się, że osoby prowadzące moderację na platformach Meta, oceniając podobne materiały zawierające dezinformację lub próby podszywania się, decydują się na usunięcie jedynie części z nich, pozostawiając resztę bez reakcji. OAD NASK musi wielokrotnie zgłaszać te same materiały, gdyż zespół moderujący nie kieruje się jednolitymi kryteriami podczas oceny zgłoszeń.

B. Administracja często nie podejmowała adekwatnej reakcji w kontekście trwających konfliktów zbrojnych na Bliskim Wschodzie. W analizowanym okresie analitycy i analityczki OAD NASK zgłaszali wpisy publikowane na platformach Facebook, Instagram oraz Threads, które powielały przekazy związane z proirańską operacją informacyjną, noszącą znamiona działań typu CIB. Pomimo tych zgłoszeń platforma nie podejmowała działań skutecznie ograniczających rozprzestrzenianie się zgłoszonych treści. Moderacja była niewystarczająca, a liczba usuniętych materiałów znikoma.

C. Istotnym problemem pozostaje nadal podejście administracji serwisu Meta polegające na dopuszczaniu do dalszej działalności reklamodawców, którzy dopuścili się wcześniejszych naruszeń regulaminu (często wielokrotnych) oraz stosowali nieuczciwe, a niekiedy przestępcze praktyki związane z wyłudzaniem danych osobowych i środków finansowych. Takie podejście

sprzyja kontynuowaniu szkodliwych działań na platformie, a jednocześnie ogranicza skuteczność przeciwdziałania manipulacjom.

OBRAZ 36. Konto reklamodawcy wielokrotnie naruszającego regulamin platformy

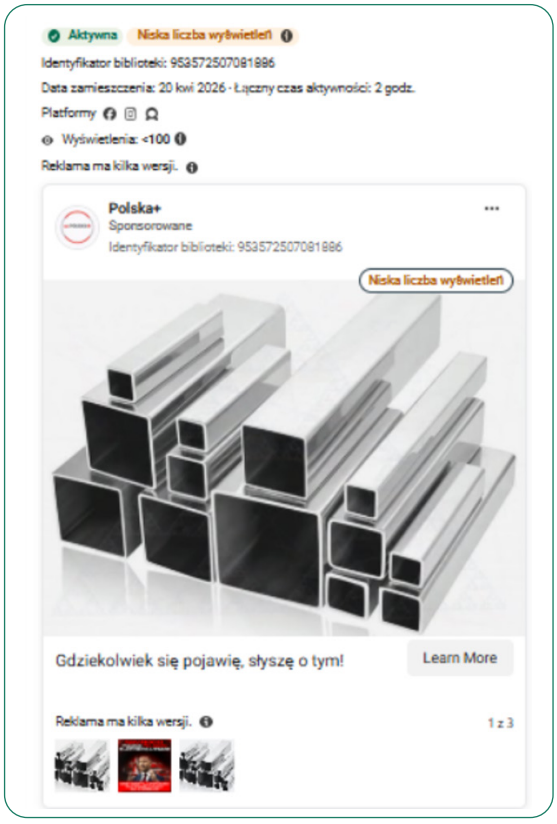


ŹRÓDŁO: FACEBOOK.COM

D. Jedną z często stosowanych przez aktorów dezinformacji metod utrudniania identyfikacji szkodliwych treści jest publikowanie tej samej reklamy w wielu wariantach (tzw. reklamy dynamiczne). Jej widoczna z poziomu biblioteki reklam wersja prowadzi zazwyczaj do przekazu niezwiązanego z faktyczną treścią, przyjmując formę neutralnego obrazu, na przykład zdjęcia przedmiotu codziennego użytku. Zdarza się również, że w wersjach udostępnianych w bibliotece reklam nie zostaje zarejestrowany pierwotny przekaz, który wcześniej był widoczny dla użytkowników i użytkowniczek. Tego rodzaju działania nie tylko utrudniają wykrycie szkodliwych treści, lecz także ograniczają możliwość przygotowania precyzyjnego zgłoszenia do administracji serwisu Meta. Opisane zjawisko może również świadczyć o nietransparentnym działaniu algorytmów na platformie, ponieważ zasady wyświetlania różnych treści odbiorcom nie są jasne ani wprost komunikowane do użytkowników.

Proceder ten był już wielokrotnie zgłaszany do punktu kontaktowego, jednak podjęte dotychczas działania nie przyczyniły się do jego skutecznego ograniczenia.

OBRAZ 37. Reklama dostępna w wielu wersjach w bibliotece reklam Meta



ŹRÓDŁO: FACEBOOK.COM

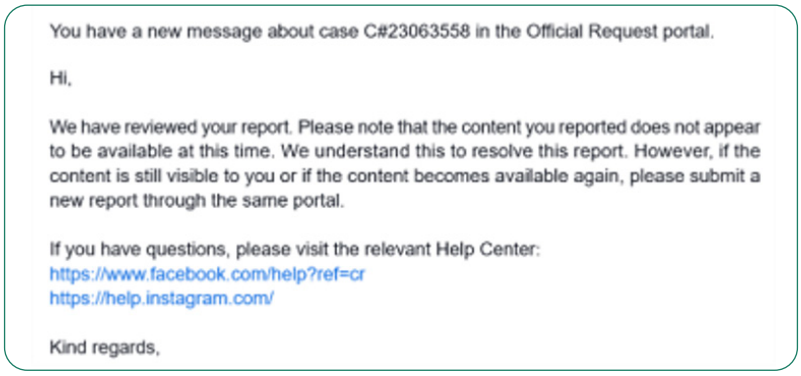
E. Utrzymującym się problemem, który znacząco utrudnia skuteczne ograniczanie zasięgu szkodliwych reklam, jest także krótki czas ich ekspozycji. Zdarza się, że administracja serwisu Meta odrzuca zgłoszenia dotyczące szkodliwych treści, uzasadniając to faktem, iż dana publikacja nie jest już dostępna dla użytkowników i użytkowniczek. Pozostawienie tego typu zgłoszeń bez odpowiedniej moderacji sprzyja dalszej aktywności nieuczciwych reklamodawców i umożliwia publikowanie kolejnych, zbliżonych przekazów. Stosowanie krótkiego czasu ekspozycji reklamy jest celowym obchodzeniem mechanizmów moderacji przez aktorów dezinformacji. Mimo wielu zgłoszeń platforma nie podejmuje w tym obszarze żadnych zmian. Przykładem niewdrożonego rozwiązania mogłoby być na przykład nakładanie na materiały reklamodawców minimalnego czasu wyświetlania. Pozwoliłoby to na wyeliminowanie wielu przypadków wprowadzania odbiorców i odbiorczyń w błąd za pomocą tej techniki.

F. Analitycy i analityczki OAD NASK nie zawsze otrzymują informację zwrotną dotyczącą statusu zgłoszonych materiałów. W konsekwencji niekiedy niemożliwe jest formalne zakończenie prowadzonej sprawy oraz przekazanie

zgłaszającemu podmiotowi lub osobie informacji o rzeczywistych działaniach podjętych przez platformę społecznościową.

G. Problem stanowią również niespójności pomiędzy informacją zwrotną a rzeczywistym statusem zgłaszanego materiału. W omawianym okresie na platformie Facebook pojawił się materiał typu deepfake przedstawiający rzekomą pracowniczkę państwowej służby zdrowia, która promowała podejrzane suplementy medyczne. Po zgłoszeniu filmu zespół OAD NASK otrzymał od administracji platformy wiadomość, że wysłany materiał nie jest już dostępny, dlatego incydent nie może zostać rozpatrzony. W rzeczywistości materiał był cały czas dostępny na platformie.

OBRAZ 38. **Komunikat platformy Meta w sprawie zgłoszonego materiału deepfake**


You have a new message about case C#23063558 in the Official Request portal.

Hi,

We have reviewed your report. Please note that the content you reported does not appear to be available at this time. We understand this to resolve this report. However, if the content is still visible to you or if the content becomes available again, please submit a new report through the same portal.

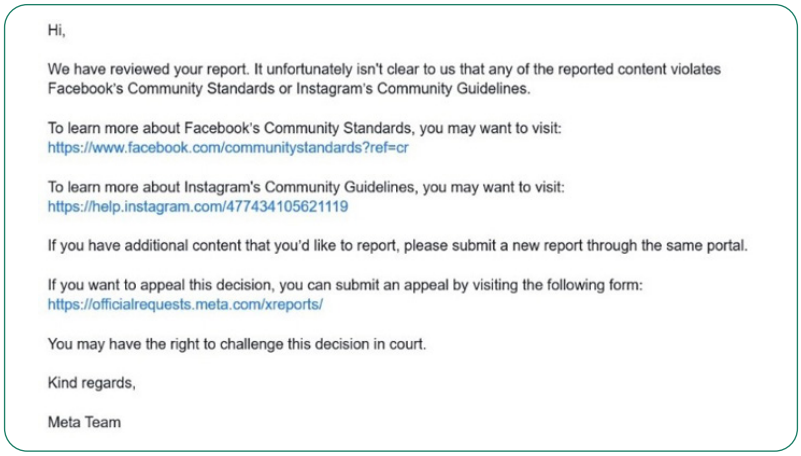
If you have questions, please visit the relevant Help Center:
https://www.facebook.com/help?ref=cr
https://help.instagram.com/

Kind regards,

ŹRÓDŁO: META OFFICIAL REQUEST PORTAL

H. Zdarza się, że decyzje moderatorów platformy Meta nie uwzględniają specyfiki krajowego kontekstu językowego, kulturowego lub społeczno-politycznego, który ma kluczowe znaczenie dla prawidłowej oceny zgłaszanych materiałów. W konsekwencji treści, które z perspektywy lokalnych uwarunkowań są uznawane za dezinformację lub manipulację, mogą zostać zaklasyfikowane jako nienaruszające standardów społeczności i pozostawać dostępne na platformie. Reakcja jest podobna bez względu na to, czy jest to pierwsze zgłoszenie, czy odwołanie.

OBRAZ 39. **Komunikat platformy Meta w sprawie dezinformacyjnego materiału na temat rzezi wołyńskiej**



Hi,

We have reviewed your report. It unfortunately isn't clear to us that any of the reported content violates Facebook's Community Standards or Instagram's Community Guidelines.

To learn more about Facebook's Community Standards, you may want to visit:
https://www.facebook.com/communitystandards?ref=cr

To learn more about Instagram's Community Guidelines, you may want to visit:
https://help.instagram.com/477434105621119

If you have additional content that you'd like to report, please submit a new report through the same portal.

If you want to appeal this decision, you can submit an appeal by visiting the following form:
https://officialrequests.meta.com/xreports/

You may have the right to challenge this decision in court.

Kind regards,

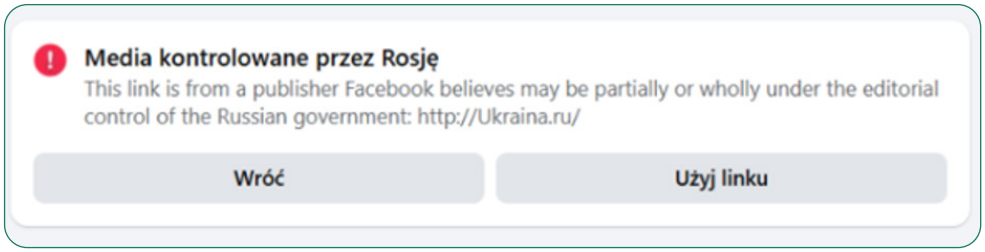
Meta Team

ŹRÓDŁO: META OFFICIAL REQUEST PORTAL

I. Odnotowano brak zadowalającej reakcji platformy na zgłoszenia związane z funkcjonowaniem grup wykorzystywanych do rozpowszechniania dezinformacji i dokonywania oszustw na odbiorcach. Zamieszczane są na nich głównie sensacyjne fake newsy, mające skłonić użytkowników i użytkowniczki do kliknięcia w link prowadzący do stron wyłudzających dane oraz pieniądze. Choć treści dezinformacyjne stanowią większość materiałów publikowanych na tych grupach, administracja serwisu nie zdecydowała się na jej usunięcie, nie podając przy tym przyczyny. Podjęto działania wyłącznie wobec pojedynczych postów, które zostały zablokowane, pozostawiając bez reakcji całą grupę, gdzie treści dezinformacyjne są publikowane masowo i regularnie.

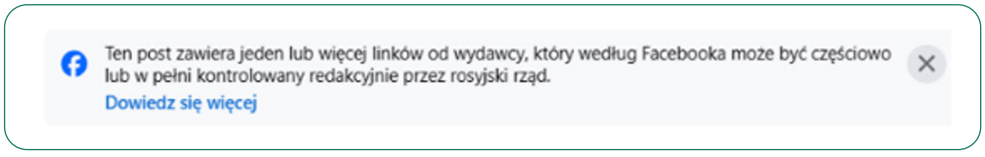
J. W omawianym okresie zaobserwowano pojawienie się na platformie Facebook nowej adnotacji na oznaczenie materiałów, które pochodzą z mediów kontrolowanych przez państwo. W przypadku przedstawionym na obrazie był to wpis, który powoływał się na propagandowy, prorosyjski portal Ukraina.ru.

OBRAZ 40. Adnotacja dotycząca mediów zależnych od państwa



ŹRÓDŁO: FACEBOOK.COM

OBRAZ 41. Adnotacja dotycząca mediów zależnych od państwa



ŹRÓDŁO: FACEBOOK.COM

Jak tłumaczy Meta, oznaczenie to dodawane jest do mediów, wobec których zachodzi podejrzenie, że nie są w pełni niezależne. Administracja platformy dokonuje oceny takich treści na podstawie wewnętrznych kategorii i standardów. Choć tego rodzaju oznaczenie może wskazywać użytkownikom i użytkowniczkom treści potencjalnie dezinformacyjne, to jednak nie ogranicza samego dostępu do dezinformacji.

OBRAZ 42. Opis adnotacji „Media kontrolowane przez państwo”



ŹRÓDŁO: FACEBOOK.COM

Podsumowanie: współpraca z platformami – trudności i wyzwania

TikTok

Proces realizacji zgłoszeń przekazywanych administracji platformy TikTok jest związany z szeregiem problemów i wyzwań.

A. Analitycy i analityczki OAD NASK regularnie odnotowują przypadki wskazujące na niekonsekwentne podejście administracji serwisu TikTok do zgłaszanych incydentów. Zgromadzone dane sugerują, że zjawisko to jest szczególnie widoczne w sprawach dotyczących podszywania się pod osoby publiczne. Choć część kont wykorzystujących wizerunek takich osób została usunięta przez zespół moderujący, w serwisie nadal funkcjonują profile podszywające się między innymi pod premiera Donalda Tuska.

B. System raportowania udostępniony przez platformę TikTok nie oferuje zaawansowanych funkcji filtrowania ani sortowania zgłoszeń. Brakuje również możliwości wyszukiwania zgłoszonych treści według autora czy bezpośredniego odnośnika do publikacji. W rezultacie czego powrót do konkretnych materiałów oraz weryfikacja ich statusu pozostają procesami czasochłonnymi i mało efektywnymi.

C. Wśród opcji dostępnych na dedykowanej platformie nie ma możliwości zgłaszania reklam na TikToku. W omawianym okresie zespół OAD NASK zidentyfikował wprowadzające w błąd reklamy wykorzystujące wygenerowany przez AI wizerunek prezesa NBP Adama Glapińskiego oraz dziennikarki Moniki Olejnik. Reklamy przedstawiały sensacyjny nagłówek pochodzący rzekomo z portalu Gazeta.pl i przekierowywały do strony imitującej tenże. Ich celem było wyłudzanie danych oraz środków pieniężnych.

Reklamy widoczne w bibliotece reklam mogą być zgłaszane wyłącznie poprzez punkt kontaktowy, a nie bezpośrednio do administracji, jak ma to miejsce w przypadku raportowania szkodliwych publikacji za pomocą platformy zgłoszeniowej. Może to wpłynąć na wydłużenie czasu oczekiwania na reakcję, przez co szkodliwy materiał będzie dłużej dostępny.

Platformy Google (YouTube, Google Ads, Google Play)

Proces realizacji zgłoszeń dezinformacji na platformach należących do Google jest związany z szeregiem problemów i wyzwań.

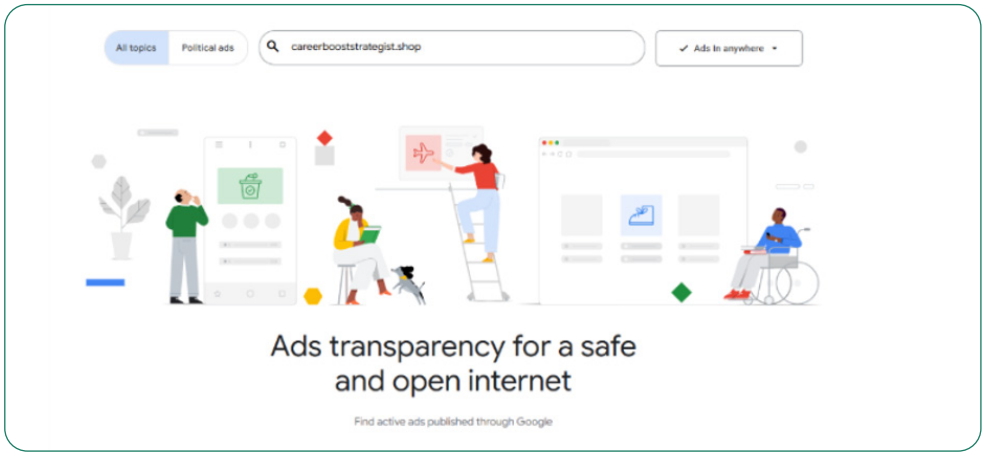
A. W związku z ograniczeniem możliwości zgłaszania szkodliwych treści za pośrednictwem formularzy kontaktowych wyłącznie do dwóch usług, YouTube oraz Google Play, monitorowanie aktualnego statusu zgłoszeń

pozostaje utrudnione. Osoba wyznaczona przez Google do kontaktu pełni jedynie funkcję pośrednika, przekazując zgłoszenia do właściwych zespołów moderujących. W konsekwencji analitycy i analityczki OAD NASK tylko sporadycznie otrzymują informacje zwrotne dotyczące działań podjętych przez serwis.

B. Biblioteka reklam udostępniona przez Google umożliwia przeszukiwanie treści wyłącznie na podstawie nazwy reklamodawcy lub odnośnika zawartego w publikacji. Brak funkcjonalności wyszukiwania według słów kluczowych istotnie ogranicza możliwość identyfikacji treści jako elementów szerszych kampanii dezinformacyjnych.

Jednocześnie, pomimo dostępności opcji wyszukiwania na podstawie linku, narzędzie to nie zawsze zapewnia skuteczną identyfikację konkretnych materiałów, co dodatkowo utrudnia prowadzenie analiz.

OBRAZ 43. **Zrzut ekranu przedstawiający próbę wyszukania reklamy po linku**



ŹRÓDŁO: GOOGLE.COM

C. Administracja platformy podejmuje niekonsekwentne i nieskuteczne działania w odpowiedzi na zgłoszenia materiałów AI. Google najczęściej reaguje na zgłoszenia pojedynczych treści stworzonych za pomocą narzędzi generatywnych. Rzadziej podejmuje działania w odniesieniu do kont, które w całości opierają się na tego rodzaju materiałach.

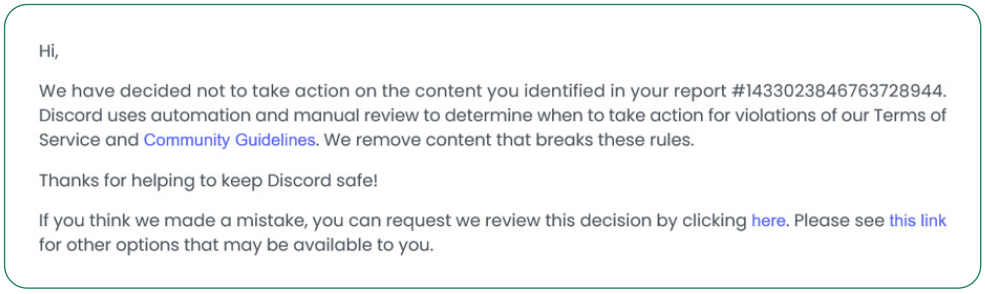
D. Zachodzą nieścisłości pomiędzy podglądem zawartości danego kanału na platformie YouTube a materiałami, które rzeczywiście udostępnia. Zaobserwowano między innymi dezinformacyjny materiał deepfake, który dostępny jest pod konkretnym linkiem, jednak nie można go znaleźć na kanale, który go opublikował. Może to być rezultat działań moderacyjnych platformy, jednak ze strony administracji nie pojawiła się odpowiednia informacja na ten temat.

Discord

Proces realizacji zgłoszeń przekazywanych administracji platformy Discord jest procesem problematycznym.

W systemie dostępnym z poziomu użytkownika po wysłaniu zgłoszenia przez formularz DSA osoba zgłaszająca otrzymuje mail z potwierdzeniem oraz numerem referencyjnym. Następnie po rozpatrzeniu sprawy platforma wysyła wiadomość ze swoją decyzją. W przypadku decyzji odmownej, funkcja odwołania nie działa.

OBRAZ 44. **Wiadomość od administracji platformy Discord w sprawie zgłoszenia**


Hi,

We have decided not to take action on the content you identified in your report #1433023846763728944. Discord uses automation and manual review to determine when to take action for violations of our Terms of Service and Community Guidelines. We remove content that breaks these rules.

Thanks for helping to keep Discord safe!

If you think we made a mistake, you can request we review this decision by clicking here. Please see this link for other options that may be available to you.

ŹRÓDŁO: POTWIERDZENIE E-MAIL Z PLATFORMY DISCORD


Ośrodek Analizy
Dezinformacji
NASK