Post w języku polskim wycelowany przeciwko społeczności osób transpłciowych

2023-023-FB-UA

Streszczenie

Rada Oversight Board uchyliła pierwotną decyzję Mety o pozostawieniu na Facebooku postu, w którym jeden z użytkowników atakował osoby transpłciowe i wypowiadał się w brutalny sposób, namawiając członków tej społeczności do popełnienia samobójstwa. Rada uznaje, że post naruszył standardy społeczności dotyczące mowy nienawiści oraz samobójstw i samookaleczeń. Jednak zasadniczą kwestią w tym przypadku nie są same zasady, ale ich egzekwowanie. Powtarzające się zaniechanie przez Metę podjęcia stosownych działań w zakresie egzekwowania zasad pomimo licznych sygnałów o szkodliwej treści postu pozwala radzie wnioskować, że spółka nie wywiązuje się z własnych deklaracji w zakresie bezpieczeństwa osób ze społeczności LGBTQIA+. Rada niniejszym wzywa Metę do staranniejszego egzekwowania zasad, również poprzez modyfikację wewnętrznych wytycznych dla osób sprawdzających treści.

Opis sprawy

W kwietniu 2023 roku użytkownik serwisu Facebook w Polsce opublikował zdjęcie przedstawiające zasłonę w paski w kolorach niebieskim, różowym i białym, czyli kolorach flagi społeczności osób transpłciowych, i napisem w języku polskim: „Nowa technologia... Zasłony, które wieszają się same". Nad zdjęciem widniały słowa „wiosenne porządki <3." Opis użytkownika zawiera słowa „Jestem transfobem". Post miał poniżej 50 reakcji.

Między kwietniem a majem 2023 r. 11 różnych użytkowników zgłosiło post łącznie 12 razy. Tylko dwa z dwunastu zgłoszeń zostały potraktowane przez zautomatyzowane systemy Meta priorytetowo jako konieczne do sprawdzenia przez człowieka, a pozostałe zostały zamknięte. Te dwa zgłoszenia przesłane do sprawdzenia przez człowieka pod kątem potencjalnego naruszenia standardów Facebooka dotyczących samobójstw i samookaleczeń zostały ocenione jako nienaruszające tychże standardów. Żadne ze zgłoszeń dotyczących mowy nienawiści nie zostało przesłane do sprawdzenia przez człowieka.

Następnie troje użytkowników odwołało się od decyzji Meta o pozostawieniu postu na Facebooku, a jedno odwołanie spowodowało, że osoba sprawdzająca treść podtrzymała

pierwotną decyzję opartą na standardach społeczności dotyczących samobójstw i samookaleczeń. Pozostałe odwołania złożone na podstawie standardu społeczności dotyczącego mowy nienawiści nie zostały przesłane do sprawdzenia przez człowieka. Jeden z użytkowników, który pierwotnie zgłosił tę treść, odwołał się ostatecznie do rady Oversight Board. Po zajęciu się tą sprawą przez radę Meta uznała, że treść faktycznie naruszyła zasady przeciwdziałania mowie nienawiści oraz samobójstwom i samookaleczeniom, i usunęła post z Facebooka. Ponadto spółka zablokowała konto użytkownika, który opublikował te treści, z powodu kilku wcześniejszych naruszeń standardów.

Najważniejsze ustalenia

Rada uznała, że opublikowane treści naruszają zasady Mety dotyczące mowy nienawiści, ponieważ zawierają „mowę nienawiści" w postaci nawoływania do samobójstwa osób z grupy podlegającej ochronie. Post nawołujący do samobójstwa osób transpłciowych stwarza atmosferę zastraszania i wykluczenia oraz może przyczynić się do powstania uszczerbku na zdrowiu i życiu. Biorąc pod uwagę sam charakter tekstu i obrazu, post przyczynia się także do pogłębiania kryzysu w zakresie zdrowia psychicznego, jakiego doświadcza społeczność osób transpłciowych. W opublikowanym niedawno raporcie autorstwa organizacji Gay and Lesbian Alliance Against Defamation (GLAAD) stwierdzono, że w Internecie widoczny jest „traumatyczny wpływ związany z ciągłym narażeniem na obelgi i postępowanie nacechowane nienawiścią". Zdaniem rady Oversight Board na poparcie jej wniosków można wskazać szerszy kontekst problemów, z jakimi boryka się społeczność osób LGBTQIA+ w Polsce w Internecie i poza nim, w tym ataki i retorykę ze strony wpływowych osobistości zajmujących stanowiska rządowe i publiczne.

Rada wyraża zaniepokojenie faktem, że osoby sprawdzające treści nie zwróciły uwagi na kontekst zgłoszenia. Odniesienie do zwiększonego ryzyka samobójstwa („zasłony, które same się wieszają") i życzenia śmierci dla społeczności („wiosenne porządki") stanowiły jawne naruszenie standardu społeczności dotyczącego mowy nienawiści, natomiast samookreślenie się twórcy treści jako „transfoba" samo w sobie było kolejnym naruszeniem standardów. Rada nalega, aby Meta poprawiła skuteczność egzekwowania zasad w zakresie nawoływania do nienawiści wobec osób LGBTQIA+, zwłaszcza gdy posty zawierają obrazy i tekst, których interpretacja wymaga informacji kontekstowych. W tym przypadku swego rodzaju zakodowane odniesienia do samobójstwa w połączeniu z wizualnym przedstawieniem grupy podlegającej ochronie (flaga społeczności osób transpłciowych) przyjęły formę „złośliwej kreatywności". Określenie to odnosi się działających w złej wierze osób opracowujących nowe sposoby atakowania społeczności osób LGBTQIA+ za pośrednictwem postów i memów przedstawianych jako „humorystyczne lub satyryczne", które jednak w rzeczywistości stanowią przejaw mowy nienawiści lub nękania.

Ponadto rada jest zaniepokojona stwierdzeniem Mety, że nieusunięcie treści przez osoby sprawdzające treści jest równoznaczne ze ścisłym stosowaniem wewnętrznych wytycznych. Oznaczałoby to, że wewnętrzne wytyczne Mety w niedostatecznym stopniu odzwierciedlają sposoby, w jakie połączenie tekstu i obrazu może przedstawiać grupę zdefiniowaną na podstawie tożsamości płciowej jej członków.

Chociaż omawiany post wyraźnie narusza standard społeczności Facebooka dotyczący samobójstw i samookaleczeń, rada uważa także, że ta zasada powinna w wyraźniejszy sposób zakazywać publikowania treści promujących samobójstwo skierowanych do możliwej do zidentyfikowania grupy osób, a nie tylko do konkretnej osoby należącej do tej grupy.

W tym przypadku systemy automatycznego ustalania priorytetów sprawdzania treści wdrożone przez Metę znacząco wpłynęły na egzekwowanie zasad, z czym wiąże się także podejście spółki do treści zgłaszanych przez wielu użytkowników. Meta monitoruje i usuwa powtarzające się zgłoszenia, aby zapewnić spójność decyzji osób sprawdzających treści i działań następczych. Inne powody automatycznego zamykania zgłoszeń obejmowały niską wagę i ograniczone rozpowszechnianie („liczbę wyświetleń, jakie zgromadziła treść"), co oznaczało, że nie były one traktowane priorytetowo jako konieczne do sprawdzenia przez człowieka. W tym przypadku rada uznała, że jako jeden z istotnych sygnałów przy ustalaniu wagi zgłoszenia należało uwzględnić opis użytkownika.

Rada uważa, że Meta powinna zająć się bardziej intensywnie opracowaniem klasyfikatorów, które pozwalają identyfikować treści potencjalnie naruszające standardy i mające wpływ na społeczność osób LGBTQIA+, a także poprawić jakość szkoleń dla osób sprawdzających treści pod kątem kwestii dotyczących naruszeń związanych z identyfikacją płciową.

Postanowienie rady Oversight Board

Rada Oversight Board uchyla pierwotną decyzję Mety o pozostawieniu postu.

Rada zaleca, aby Meta:

- Na stronie standardów dotyczących samobójstw i samookaleczeń dodała, że zasada zabrania publikowania treści promujących samobójstwo lub zachęcających do samobójstwa i skierowanych do możliwej do zidentyfikowania grupy osób.
- Zmodyfikowała wewnętrzne wytyczne przekazywane osobom sprawdzającym masowo zgłaszane treści, tak aby upewnić się, że wizualne przedstawienia dotyczące tożsamości płciowej, które nie zawierają postaci ludzkiej, ale obejmują flagi, były rozumiane jako przedstawienie grupy określonej na podstawie tożsamości płciowej jej członków.

*Streszczenia stanowią ogólny opis spraw i nie stanowią punktu odniesienia.

Postanowienie rady

1. Streszczenie postanowienia

1. Rada Oversight Board uchyla niniejszym pierwotną decyzję Mety o pozostawieniu na Facebooku treści w języku polskim, w której to treści jeden z użytkowników atakował osoby transpłciowe i wypowiadał się w brutalny sposób, namawiając członków tej społeczności do popełnienia samobójstwa. Po skierowaniu przez radę sprawy do ponownego rozpatrzenia Meta stwierdziła, że pierwotna decyzja zezwalająca na pozostawienie wpisu na platformie była błędna, a następnie usunęła treść i nałożyła sankcje na użytkownika. Rada uznała, że post naruszył standardy społeczności dotyczące mowy nienawiści oraz samobójstw i samookaleczeń. Rada pragnie przy okazji Mecie ulepszenie zasad i wytycznych dla osób sprawdzających treści, aby skuteczniej chronić osoby transpłciowe na jej platformach. Modyfikacje te są konieczne, by Meta zapewniła, że posty zgłaszane jako potencjalna mowa nienawiści i zawierające wizualne przedstawienia dotyczące tożsamości płciowej, które nie zawierają postaci ludzkiej, ale obejmują flagi, będą rozumiane jako przedstawienie grupy określonej na podstawie tożsamości płciowej jej członków. Meta powinna również wyjaśnić, że zachęcanie całej grupy do popełnienia samobójstwa jest takim samym naruszeniem, jak zachęcanie do popełnienia samobójstwa pojedynczej osoby. Rada stwierdza jednak, że zasadniczą kwestią w tym przypadku nie są same zasady, ale ich egzekwowanie. Zasada w bieżącym brzmieniu wyraźnie zabraniała opublikowania postu, który zawierałby wiele przykładów mowy nienawiści skierowanej do grupy osób możliwej do zidentyfikowania ze względu na tożsamość płciową jej członków. Powtarzające się zaniechanie przez Metę podjęcia w tym przypadku stosownych działań w zakresie egzekwowania zasad pomimo licznych zgłoszeń użytkowników pozwala radzie wnioskować, że Meta nie wywiązuje się z własnych deklaracji w zakresie bezpieczeństwa osób ze społeczności LGBTQIA+. Rada niniejszym wzywa Metę do staranniejszego egzekwowania zasad.

2. Opis i szczegółowe informacje dotyczące sprawy

2. W kwietniu 2023 roku pewien użytkownik serwisu Facebook w Polsce opublikował zdjęcie przedstawiające zasłonę w paski w kolorach niebieskim, różowym i białym, czyli kolorach flagi

osób transpłciowych. Na obrazie umieszczono następujący tekst w języku polskim: „Nowa technologia. Zasłony, które wieszają się same". Nad tym tekstem znalazły się słowa w języku polskim: „wiosenne porządki <3." Opis użytkownika w języku polskim zawiera słowa „Jestem transfobem". Post miał poniżej 50 reakcji, z czego większość była przychylna treści. Najczęściej używanym emotikonem był emotikon śmiechu, „Haha".

3. W okresie od kwietnia do maja 2023 roku 11 różnych użytkowników zgłosiło tę treść, a liczba zgłoszeń wyniosła łącznie 12. Spośród tych zgłoszeń dziesięć nie zostało potraktowanych przez zautomatyzowane systemy Meta priorytetowo jako konieczne do sprawdzenia przez człowieka z różnych powodów, w tym z powodu „niskich wskaźników wagi i rozpowszechniania". Meta ogólnie ustanawia priorytety dotyczące sprawdzania treści przez człowieka na podstawie wagi, wskaźnika rozpowszechniania i prawdopodobieństwa naruszenia zasad dotyczących treści. Tylko dwa zgłoszenia, dotyczące standardu społeczności Facebooka dotyczącego samobójstw i samookaleczeń, zostały przesłane do sprawdzenia przez człowieka. Żadne ze zgłoszeń dotyczących zasady zapobiegania mowie nienawiści nie zostało przesłane do sprawdzenia przez człowieka. Według Mety osoby sprawdzające treści są przeszkolone w zakresie „oceny i podejmowania działań" (oraz mają stosowne narzędzia do podejmowania takich dziań) pod kątem treści wykraczających poza przypisane im zasady (np. dotyczące mowy nienawiści lub samobójstw i samookaleczeń). Niemniej jednak obie osoby sprawdzające treści oceniły tę treść jako nienaruszającą zasad i nie eskalowały problemu.

4. Troje użytkowników złożyło odwołanie od decyzji Mety dotyczącej pozostawienia tej treści na Facebooku. W odpowiedzi na jedno odwołanie osoba sprawdzająca treść podtrzymała pierwotną decyzję Mety mówiącą, że treść ta nie narusza polityki dotyczącej przeciwdziałania samobójstwom i samookaleczeniom. Dwa pozostałe odwołania opierające się na polityce przeciwdziałania mowie nienawiści Facebooka nie zostały przekazane do sprawdzenia przez ludzi. Wynika to z faktu, że Meta „monitoruje i usuwa powtarzające się" zgłoszenia tych samych treści, aby zapewnić spójność decyzji osób sprawdzających treści i działań następczych.

5. Jeden z użytkowników, który pierwotnie zgłosił tę treść, odwołał się następnie do rady Oversight Board. Po zajęciu się tą sprawą przez radę Meta uznała, że treść faktycznie naruszyła zasady przeciwdziałania mowie nienawiści oraz samobójstwom i samookaleczeniom, i usunęła post. Ponadto w ramach weryfikacji sprawy przez Metę ustalono, że na koncie twórcy tej treści doszło już do kilku naruszeń standardów społeczności i osiągnięto próg naruszeń pozwalający na zablokowanie konta. Meta zablokowała to konto w sierpniu 2023 roku.

6. Podejmując decyzję w tej sprawie, rada odniosła się do następującego kontekstu:

7. W Polsce często zgłaszany jest wysoki poziom wrogości wobec społeczności osób LGBTQIA+. [Uwaga: rada używa określenia „LGBTQIA+" (lesbijki, geje, osoby biseksualne, transpłciowe, queer, interseksualne i aseksualne) w odniesieniu do grup identyfikowanych w oparciu o koncepcje orientacji seksualnej, tożsamości płciowej lub ekspresji płciowej. Jednakże rada będzie stosować akronimy lub zwyczajowe określenia stosowane przez inne osoby lub podmioty podczas cytowania lub przytaczania ich wypowiedzi]. Komisarz Praw Człowieka Rady Europy wcześniej zwrócił uwagę na problem „stygmatyzacji osób LGBTI" jako „długoletni problem w Polsce". Raport Międzynarodowego Stowarzyszenia Lesbijek, Gejów, Osób Biseksualnych, Transpłciowych i Interseksualnych (ILGA) i oddziału Rainbow Europe klasyfikuje kraje na podstawie przepisów i polityk, które bezpośrednio wpływają na prawa człowieka osób LGBTI. W raporcie tym Polska jako państwo członkowskie Unii Europejskiej (UE) osiągnęła najniższe wyniki i zajmuje 42. miejsce na 49 ocenionych krajów europejskich. Władze krajowe i lokalne, a także znaczące osobistości w coraz większym stopniu używają wobec społeczności LGBTQIA+ dyskryminujących określeń w przemowach, a także podejmują dyskryminujące działania legislacyjne.

8. Począwszy od 2018 r. organizacja ILGA-Europe monitoruje kwestię, którą nazywa „otwartą polityczną mową nienawiści wobec osób LGBTI ze strony polskich przywódców politycznych", w tym twierdzenia stanowiące, że „cały ruch LGBT" stanowi „zagrożenie" dla Polski. W tym samym roku prezydent Lublina podjął próbę zakazania Marszu Równości, ale Sąd Apelacyjny uchylił zakaz na krótko przed planowanym marszem. W 2019 roku Prezydent miasta stołecznego Warszawy wprowadził Warszawską Politykę Miejską Na Rzecz Społeczności LGBT+ (Deklarację LGBT+) mającą na celu „poprawę sytuacji społeczności LGBT" w mieście. Rządząca w Polsce partia Prawo i Sprawiedliwość (PiS) oraz przywódcy religijni skrytykowali tę deklarację. Prezydent Polski i rząd wypowiadali się również krytycznie na temat społeczności osób transpłciowych. Na przykład prezes partii rządzącej, PiS, określił osoby transpłciowe jako „nienormalne". Minister Sprawiedliwości zwrócił się także do Sądu Najwyższego o rozważenie postanowienia, na mocy którego „oprócz rodziców osoby transpłciowe powinny także pozywać swoje dzieci i współmałżonka [o pozwolenie na tranzycję], jeśli chcą uzyskać dostęp do LGR [Legal Gender Recognition, prawnej korekty płci]".

9. Polska przyjęła również przepisy anty-LGBTQIA+. Według organizacji Human Rights Watch polskie miejscowości rozpoczęły akcje na rzecz „wykluczenia osób LGBT z polskiego społeczeństwa" poprzez tworzenie, między innymi, „stref wolnych od LGBT" w 2019 r. Organizacja Human Rights Watch podała, że strefy te to miejsca, „w których władze lokalne przyjęły dyskryminujące »karty rodziny«, zobowiązując się do »ochrony dzieci przed zepsuciem moralnym« lub ogłosiły, że są wolne od »ideologii LGBT«". Strefy takie utworzyło ponad 100 miejscowości. ILGA-Europe informuje, że pod presją lokalną, unijną i międzynarodową niektóre z tych gmin wycofały „uchwały anty-LGBT lub Karty praw rodziny". W dniu 28 czerwca 2022 r. Naczelny Sąd Administracyjny wydał nakaz wycofania uchwał skierowanych przeciwko LGBTQI+ w czterech gminach. Niemniej jednak, jak sugeruje ranking znajdujący się w raporcie Rainbow Europe, klimat społeczny w Polsce jest szczególnie nieprzyjazny społeczności osób LGBTQIA+.

10. Przeprowadzone przez Agencję Praw Podstawowych Unii Europejskiej w roku 2019 badanie ankietowe osób LGBTI miało na celu porównanie doświadczeń osób LGBTI związanych z napaścią i nękaniem w Polsce i innych krajach Unii Europejskiej. Według ankiety 51% osób LGBTI w Polsce często lub zawsze unika pewnych miejsc w obawie przed napaścią. Dla porównania w pozostałych krajach Unii Europejskiej odsetek ten wynosi 33%. W badaniu wykazano również, że 1 na 5 osób transpłciowych doświadczyła przemocy fizycznej lub seksualnej w ciągu pięciu lat poprzedzających badanie, czyli było ich ponad dwukrotnie więcej niż w przypadku innych grup LGBTI.

11. Rada zleciła zewnętrznym ekspertom analizę reakcji w mediach społecznościowych na obraźliwe wypowiedzi polskich urzędników państwowych. Eksperci ci zauważyli „niepokojący wzrost w Internecie mowy nienawiści skierowanej przeciwko społecznościom mniejszościowym w Polsce, w tym społeczności osób LGBTQIA+, który nastąpił od 2015 roku". W analizie treści skierowanych przeciwko społeczności osób LGBTQIA+ w języku polskim na Facebooku eksperci stwierdzili, że wyraźny wzrost wystąpił po publikacji „orzeczeń sądowych odnoszących się do ustawodawstwa przeciwdziałającego LGBTQIA+". Należą do nich omawiane powyżej orzeczenie Naczelnego Sądu Administracyjnego oraz ustalenia dotyczące skarg prawnych w związku z przyjęciem deklaracji anty-LGBT wniesionych do lokalnych sądów administracyjnych przez polskich rzeczników z biura Rzecznika Praw Obywatelskich, które są rozpatrywane od 2019 roku.

12. Rada zwróciła się także do lingwistów o wyjaśnienie znaczenia dwóch polskich zwrotów zawartych w poście. W odniesieniu do wyrażenia „zasłony, które wieszają się same" eksperci stwierdzili, że w kontekście „flagi społeczności osób transpłciowych wiszącej w oknie" wyrażenie to stanowiło „grę słów", w której zestawiono „powieszenie zasłon" z „popełnieniem samobójstwa przez powieszenie". Eksperci doszli do wniosku, że sformułowanie to było „zawoalowaną transfobiczną obelgą". Jeśli chodzi o wyrażenie „wiosenne porządki", eksperci stwierdzili, że wyrażenie to „zwykle odnosi się do dokładnego sprzątania domu na wiosnę", ale w niektórych kontekstach „oznacza także „wyrzucenie wszystkich śmieci" i „pozbycie się wszystkich niechcianych przedmiotów (lub osób)". W kilku publicznych komentarzach, w tym w komentarzach fundacji Human Rights Campaign (PC-16029), stwierdzono, że zawarte w poście odniesienie do „wiosennych porządków" było formą „pochwały wykluczenia i izolacji osób transpłciowych ze społeczeństwa polskiego (poprzez śmierć)".

13. W tym przypadku kwestie szkodliwości w Internecie i poza nim wykraczają poza społeczność LGBTQIA+ w Polsce i wpływają na tę społeczność na całym świecie. Według Światowej Organizacji Zdrowia (WHO) samobójstwo jest czwartą najczęstszą przyczyną śmierci wśród osób w wieku 15–29 lat na całym świecie. WHO podaje, że „wskaźniki samobójstw są również wysokie wśród grup szczególnie wrażliwych, które doświadczają dyskryminacji, takich jak uchodźcy i migranci, ludność tubylcza oraz lesbijki, geje, osoby biseksualne, transpłciowe i interseksualne (LGBTI). W badaniach naukowych wykazano „dodatnią korelację" pomiędzy cyberprzemocą a myślami i zachowaniami prowadzącymi do samookaleczeń.

14. Ryzyko samobójstwa dotyczy szczególnie społeczności osób transpłciowych i niebinarnych. W ramach projektu ankietowego Trevor 2023 National Survey dotyczącego zdrowia psychicznego społeczności LGBTQ wykazano, że połowa młodych osób transpłciowych i niebinarnych w Stanach Zjednoczonych rozważała próbę samobójczą w 2022 roku. Według szacunków opartych na tym samym badaniu 14% młodych osób LGBTQ podjęło próbę samobójczą w ciągu ostatniego roku, z czego próbę taką podjęło prawie 1 na 5 młodych osób transpłciowych i niebinarnych. Według badania zachowań ryzykownych wśród młodzieży przeprowadzonego przez CDC (Youth Risk Behavior Survey) w 2021 roku 10% uczniów szkół średnich w Stanach Zjednoczonych próbowało popełnić samobójstwo. Liczne badania prowadzone na całym świecie potwierdzają, że osoby transpłciowe i niebinarne są bardziej narażone na myśli i próby samobójcze w porównaniu z osobami cispłciowymi.

15. W publicznym komentarzu skierowanym do rady stowarzyszenie Gay and Lesbian Alliance Against Defamation (GLAAD) (PC-16027) podkreśliło wnioski z corocznego badania ankietowego Social Media Safety Index dotyczącego bezpieczeństwa mediów społecznościowych w kwestii bezpieczeństwa użytkowników należących do społeczności LGBTQ na pięciu głównych platformach mediów społecznościowych. W raporcie za rok 2023 wynik portalu Facebook na podstawie 12 wskaźników specyficznych dla LGBTQ wyniósł 61%. Wynik ten oznacza wzrost o 15 punktów procentowych w porównaniu z rokiem 2022 i oznacza, że Facebook zajął drugie miejsce po Instagramie i wyprzedził trzy inne główne platformy. Jednak GLAAD dodaje: „bezpieczeństwo i jakość ochrony użytkowników ze społeczności LGBTQ pozostają niezadowalające". W raporcie stwierdzono, że „w Internecie osobom ze społeczności LGBTQI wyrządza się realne szkody, włącznie z wywieraniem mrożącego wpływu na wolność wyrażania opinii przez osoby LGBTQ, które boją się stać celem przemocy, oraz czysto traumatycznymi skutkami psychologicznymi wynikającymi z trwałego narażenia na obelgi i zachowanie nacechowane nienawiścią".

3. Uprawnienia i zakres działania rady Oversight Board

16. Rada ma prawo dokonać weryfikacji decyzji Mety w następstwie odwołania złożonego przez osobę, która wcześniej zgłosiła treść nieusuniętą z platformy (art. 2 Statutu, część 1; art. 3 Przepisów wewnętrznych, część 1).

17. Rada może podtrzymać lub uchylić decyzję Mety (art. 3 Statutu, część 5), a postanowienie to jest wiążące dla spółki (art. 4 Statutu). Meta musi także ocenić wykonalność decyzji w odniesieniu do identycznych treści w równoległym kontekście (art. 4 Statutu). Postanowienia rady mogą zawierać niewiążące zalecenia, na które Meta musi zareagować (art. 3 Statutu, część 4; art. 4). Jeżeli Meta zobowiąże się do działania zgodnie z zaleceniami, rada będzie monitorować realizację tych działań.

18. W przypadkach, gdy rada zajmuje się takimi przypadkami, jak rozważany, w których Meta przyznaje, że popełniono błąd, rada dokonuje weryfikacji pierwotnej decyzji w celu poprawy zrozumienia procesów moderowania treści i wydania zaleceń w zakresie ograniczania błędów i bardziej uczciwego traktowania osób korzystających z Facebooka i Instagrama.

4. Źródła uprawnień i wytycznych

19. Na analizę rady w tej sprawie miały wpływ następujące standardy i precedensy:

I. Postanowienia rady Oversight Board:

20. Do najistotniejszych wcześniejszych postanowień rady Oversight Board uwzględnionych w tej sprawie należą:

- Przywrócenie postu ze słowami w języku arabskim
- Rysunek miasta Knin
- Protesty w Kolumbii
- Ormianie w Azerbejdżanie

II. Zasady Mety dotyczące publikowanych treści:

21. Uzasadnienie zasady dotyczącej mowy nienawiści definiuje mowę nienawiści „jako bezpośredni atak na osoby – a nie na koncepcje czy instytucje – na podstawie . . cech chronionych", w tym płci i tożsamości płciowej. Meta definiuje „ataki" jako „brutalne lub odczłowieczające treści, szkodliwe stereotypy, oświadczenia o niższości, treści wyrażające pogardę, wstręt lub odrzucenie, przekleństwa i wezwania do wykluczenia lub segregacji". W uzasadnieniu zasady Meta stwierdza dalej: „Jesteśmy przekonani, że ludzie swobodniej zabierają głos i nawiązują kontakty, gdy nie czują się atakowani za to, kim są. Z tego względu nie zezwalamy na mowę nienawiści na Facebooku. Powoduje ona sytuację zastraszenia i wykluczenia, a w niektórych przypadkach może promować przemoc poza Internetem."

22. Standard społeczności Mety dotyczący mowy nienawiści dzieli ataki na „poziomy". Ataki Poziomu 1 obejmują treści wymierzone w osobę lub grupę osób ze względu na cechy chronione za pomocą „brutalnej mowy lub poparcia twierdzeń w postaci pisemnej lub wizualnej". Meta ostatecznie stwierdziła, że omawiany post naruszył ten punkt zasad. Dnia 6 grudnia 2023 r. Meta zaktualizowała standardy społeczności, aby odzwierciedlić fakt, że zakaz wyrażania przemocy wobec grup podlegających ochronie został przeniesiony do zasad dotyczących przemocy i podżegania do niej.

23. Ataki Poziomu 2 obejmują treści wymierzone w osobę lub grupę osób ze względu na cechy chronione za pomocą „wyrazów pogardy (w formie pisemnej lub wizualnej)". W standardzie społeczności dotyczącym mowy nienawiści Meta definiuje wyrazy pogardy jako „przyznanie się

do braku tolerancji ze względu na chronioną cechę" oraz „stwierdzenia, że chroniona cecha nie powinna istnieć".

24. Standard społeczności dotyczący samobójstw i samookaleczeń zabrania publikowania „wszelkich treści zachęcających do samobójstwa lub samookaleczenia, w tym treści fikcyjnych, takich jak memy lub ilustracje". Zgodnie z tą zasadą Meta usuwa „treści, które promują, zachęcają, koordynują lub zawierają instrukcje dotyczące samobójstwa lub samookaleczenia".

25. Analiza dokonana przez radę została oparta na zaangażowaniu Meta w wyrażanie opinii, co spółka określa jako „najważniejsze", oraz wyznawane przez nią wartości, takie jak bezpieczeństwo i godność.

III. Obowiązki Mety w zakresie przestrzegania praw człowieka

26. Wytyczne ONZ dotyczące biznesu i praw człowieka (UNGP) zatwierdzone przez Radę Praw Człowieka ONZ w 2011 roku ustanawiają dobrowolne ramy dotyczące obowiązków prywatnych przedsiębiorstw w zakresie praw człowieka. W 2021 roku Meta ogłosiła utworzenie Korporacyjnych zasad praw człowieka, w których potwierdziła swoje zaangażowanie na rzecz poszanowania praw człowieka zgodnie z UNGP. Analiza rady dotycząca odpowiedzialności Mety w zakresie praw człowieka w omawianej sprawie została oparta na następujących międzynarodowych standardach:

- Prawo do wolności opinii i wypowiedzi: Artykuł 19 Międzynarodowego Paktu Praw Obywatelskich i Politycznych (MPPOiP); Komentarz ogólny nr 34, Komisja Praw Człowieka, 2011; Specjalny Sprawozdawca ONZ (UNSR) ds. wolności opinii i wypowiedzi, raporty: A/HRC/38/35 (2018), A/74/486 (2019); i raport Plan Działania z Rabatu Wysokiego Komisarza ONZ ds. Praw Człowieka: A/HRC/22/17/Add.4 (2013).
- Prawo do życia: Artykuł 6 MPPOiP.
- Prawo do korzystania z najwyższego osiągalnego poziomu zdrowia fizycznego i psychicznego: Artykuł 12 Międzynarodowego paktu praw gospodarczych, społecznych i kulturalnych (ICESCR).
- Prawo do równości i niedyskryminacji: Artykuł 2, ustęp 1 i Artykuł 26 MPPOiP.

5. Zgłoszenia przesyłane przez użytkowników

27. W przesłanym do rady odwołaniu użytkownik, który zgłosił treść, stwierdził, że osoba, która zamieściła zdjęcie, nękała już wcześniej osoby transpłciowe w Internecie, a po zawieszeniu

konta na Facebooku utworzyła nowe konto. Użytkownik dodał, że pochwalanie wysokiego wskaźnika samobójstw w społeczności osób transpłciowych „nie powinno być dozwolone”.

6. Informacje przekazane przez Metę

28. Meta usunęła post na podstawie Poziomu 1 standardu społeczności dotyczącego mowy nienawiści, ponieważ treść naruszała zasady zabraniające publikowania treści skierowanych do osoby lub grupy osób ze względu na cechy chronione i zawierających „brutalną mowę lub poparcie twierdzeń w postaci pisemnej lub wizualnej”. W swoich wewnętrznych wytycznych dotyczących stosowania tych zasad Meta stwierdza, że treści należy usuwać, jeśli stanowią one „agresywne wypowiedzi w formie wezwań do działania lub oświadczeń o zamiarze wyrządzenia krzywdy, stwierdzenia życzące śmierci (także warunkowe) lub stwierdzenia nawołujące do śmierci, zachorowania lub szkody lub je popierające (w formie pisemnej lub wizualnej).”

29. W wewnętrznych wytycznych spółka opisuje również to, co uważa za wizualną reprezentację grup o cechach chronionych na obrazie lub filmie. Meta nie zezwoliła radzie na publikację bardziej szczegółowych informacji związanych z tymi wytycznymi. Zamiast tego spółka oświadczyła, że „zgodnie z zasadami dotyczącym mowy nienawiści Meta może wziąć pod uwagę elementy wizualne treści przy ustalaniu, czy treść jest skierowana do osoby lub grupy osób na podstawie ich cech chronionych”.

30. Meta stwierdziła, że wielokrotna ocena treści przez osoby oceniające skutkująca kwalifikacją postu jako nienaruszającego standardów społeczności dotyczących mowy nienawiści jest zgodna ze „ścisłym stosowaniem wewnętrznych wytycznych”. Następnie dodała: „Chociaż zasłony przypominają flagę Trans Pride, atak na pojedynczą flagę zinterpretowalibyśmy jako atak na koncepcję lub instytucję, która nie narusza naszych zasad, a nie jako atak na osobę lub grupy osób”. Jednak Meta ustaliła później, że „odniesienie do powieszenia wskazuje, że ten post atakuje grupę osób”. Ocena ta opierała się na ustaleniu, że sformułowanie „zasłony, które wieszają się same” w sposób dorozumiany odnosi się do wskaźnika samobójstw w społeczności osób transpłciowych, ponieważ zasłony przypominają flagę Trans Pride, a zasłony na zdjęciu (oraz nałożony na nie tekst) to metafora samobójstwa poprzez powieszenie. Meta zauważyła również, że „koncepcje lub instytucje nie mogą się »powiesić«, przynajmniej nie dosłownie”. Z tego powodu Meta ustaliła, że użytkownik miał na myśli „osoby transpłciowe, a nie tylko koncepcję”. Z tego powodu zdaniem Meta „ta treść

narusza zasady dotyczące mowy nienawiści, ponieważ należy ją interpretować jako oświadczenie na rzecz samobójstwa grupy o cechach chronionych”.

31. Po aktualizacji zasad dotyczących mowy nienawiści, na mocy której zakaz wyrażania przemocy wobec grup podlegających ochronie został przeniesiony do zasad dotyczących przemocy i podżegania do niej, Meta przekazała radzie, że treść nadal narusza standardy.

32. Meta poinformowała również, że stwierdzenie zawarte w opisie konta użytkownika („Jestem transfobem”) narusza Poziom 2 zasad dotyczących nawoływania do nienawiści jako „samodzielne przyznanie się do braku tolerancji ze względu na chronioną cechę”. Zdaniem Meta stwierdzenie to zostało uznane za naruszające standardy podczas weryfikacji sprawy oraz konta użytkownika przez Metę po zajęciu się sprawą przez radę. Meta stwierdziła, że to stwierdzenie pomogło zrozumieć intencje użytkownika opisane w sprawie.

33. W odpowiedzi na pytanie rady, czy treść narusza zasady dotyczące samobójstw i samookaleczeń, Meta potwierdziła, że „treść narusza zasady dotyczące samobójstw i samookaleczeń, zachęcając do samobójstwa, zgodnie z naszym ustaleniem, że treść stanowi popieranie śmierci osób z grupy podlegającej ochronie przez samobójstwo”. Meta poinformowała również, że zasady dotyczące samobójstw i samookaleczeń „nie wprowadzają rozróżnienia pomiędzy treściami promującymi samobójstwo lub do niego zachęcającymi i skierowanymi do konkretnej osoby a treściami skierowanymi do grupy osób”.

34. Rada zadała Mecie 13 pytań na piśmie. Pytania te dotyczyły podejścia Mety do moderowania treści dotyczących kwestii osób transpłciowych i LGBTQIA+; związku pomiędzy mową nienawiści a standardami społeczności dotyczącymi samobójstw i samookaleczeń; tego, w jaki sposób moderatorzy oceniają „humor” i „satyrę” podczas sprawdzania treści pod kątem naruszeń związanych z mową nienawiści; roli wskaźników „rozpowszechniania” i „wagi” w ustalaniu priorytetów przesyłania treści do sprawdzenia przez człowieka; a także tego, w jaki sposób praktyki moderowania treści Mety obejmują ustalanie priorytetów przesyłania treści do sprawdzenia przez człowieka w przypadku zgłoszeń od wielu użytkowników. Meta odpowiedziała na wszystkie 13 pytań.

7. Komentarze publiczne

35. Rada Oversight Board otrzymała 27 komentarzy publicznych dotyczących tej sprawy, w tym 19 ze Stanów Zjednoczonych i Kanady, sześć z Europy i dwa z regionu Azji i Pacyfiku oraz Oceanii. Nie obejmuje to siedmiu dodatkowych komentarzy publicznych, które były duplikatami lub zostały przesłane za zgodą na publikację, ale nie spełniały warunków określonych przez radę. Może to być związane z obraźliwym charakterem komentarza, obawami dotyczącymi prywatności użytkownika lub innymi przyczynami prawnymi. Komentarze publiczne można przesyłać do rady za zgodą lub bez zgody na publikację oraz za zgodą lub bez zgody na przypisanie im autorstwa.

36. Komentarze dotyczyły następujących kwestii: sytuacja w zakresie praw człowieka w Polsce, zwłaszcza w kontekście osób transpłciowych; bezpieczeństwo społeczności LGBTQIA+ na platformach mediów społecznościowych; związek pomiędzy przemocą online i offline w Polsce; związek pomiędzy humorem, satyrą, memami i nienawiścią/nękaniem wobec osób transpłciowych na platformach mediów społecznościowych; oraz wyzwania związane z moderowaniem treści, których interpretacja wymaga kontekstu.

37. Z publicznymi komentarzami przesłanymi w tej sprawie można zapoznać się tutaj.

8. Analiza rady Oversight Board

38. Rada zbadała, czy te treści powinny zostać usunięte, analizując zasady Mety dotyczące treści, a także jej zobowiązania i wartości w zakresie praw człowieka. Rada oceniła także konsekwencje omawianej sprawy dla szerszego podejścia Meta do zarządzania treścią.

39. Rada wybrała tę sprawę, aby ocenić staranność egzekwowania przez Metę zasad dotyczących mowy nienawiści, a także aby lepiej zrozumieć, w jaki sposób Meta traktuje treści zawierające zarówno mowę nienawiści, jak i zachęcanie do samobójstwa lub samookaleczenia.

8.1 Zgodność z zasadami Mety dotyczącymi publikowanych treści

I. Zasady dotyczące treści

Mowa nienawiści

40. Rada uznała, że treści w omawianej sprawie naruszyły standard społeczności dotyczący mowy nienawiści. Rada uznała, że post zawierał „mowę nienawiści lub popieranie nienawiści” (Poziom 1) w postaci nawoływania do samobójstwa osób z grupy podlegającej ochronie, co w oczywisty sposób stanowi naruszenie zasad dotyczących mowy nienawiści.

41. Rada zgadza się z ostatecznym wnioskiem Meta, że odniesienie do powieszenia zawarte w poście stanowi atak na grupę osób, a nie na koncepcję, ponieważ „koncepcje lub instytucje nie mogą się »powiesić«”. Rada uznaje również, że w szerszym kontekście przemocy w Internecie i poza nim, z jaką spotykają się członkowie społeczności LGBTQIA+, a w szczególności osoby transpłciowe, w Polsce, jej wniosek ma uzasadnienie. Post zawierający mowę nienawiści w celu nawoływania do samobójstwa osób transpłciowych i popierania go stwarza atmosferę zastraszania i wykluczenia oraz może przyczynić się do powstania uszczerbku na zdrowiu i życiu. Kontekst, w jakim użyto języka postu, jasno pokazuje, że miał on na celu odczłowieczenie. Biorąc pod uwagę sam charakter tekstu i obrazu, post przyczynia się także do pogłębiania trwającego kryzysu w zakresie zdrowia psychicznego, jakiego doświadcza obecnie społeczność osób transpłciowych. Według licznych badań osoby transpłciowe lub niebinarne są bardziej narażone na myśli i próby samobójcze w porównaniu z osobami cispłciowymi. Co więcej, ataki i wiktymizacja w Internecie są dodatnio skorelowane z myślami samobójczymi. W tym kontekście rada uznaje, że flaga społeczności osób transpłciowych w połączeniu z odniesieniem mówiącym o podwyższonym ryzyku samobójstwa w społeczności osób transpłciowych („zasłony, które wieszają się same”) wyraźnie wskazuje, że celem postu są osoby transpłciowe. Rada uważa także, że wyrażenie „wiosenne porządki”, po którym następuje emotikon „<3” (serce), również stanowi poparcie dla życzeń śmierci skierowanych do tej grupy. W związku z tym narusza on również zawarty w standardach społeczności dotyczących mowy nienawiści zakaz (Poziom 2) publikowania „stwierdzeń, że chroniona cecha nie powinna istnieć”.

42. Jeśli chodzi o wspomniany standard społeczności, rada jest zdania, że zasady i wewnętrzne wytyczne dotyczące ich egzekwowania mogłyby lepiej być lepiej dostosowane do „złośliwej kreatywności” w trendach treści skierowanych do grup historycznie marginalizowanych. Centrum Wilsona ukuło to sformułowanie na podstawie badań nad nękaniem ze względu na płeć i tożsamość płciową, a GLAAD również podkreśla jego znaczenie w publicznym komentarzu (PC-16027). „Złośliwa kreatywność” odnosi się do „używania zakodowanego języka, iteracyjnych, kontekstowych memów wizualnych i tekstowych oraz innych taktyk

mających na celu uniknięcie wykrycia treści na platformach mediów społecznościowych". Stosując tę koncepcję do omawianego postu, GLAAD stwierdza, że „złośliwa kreatywność" obejmuje „działające w złej wierze osoby opracowujące nowe sposoby atakowania społeczności LGBTQ" i szerzej grup szczególnie wrażliwych za pośrednictwem postów i memów, których osoby te bronią jako „humorystycznych lub satyrycznych", ale które są „co do zasady przejawami mowy nienawiści lub nękania osób LGBTQ". W szczególności „złośliwa kreatywność" przybrała w omawianej sprawie formę postu, w którym zastosowano dwa zakodowane odniesienia do samobójstwa („zasłony, które wieszają się same" i „wiosenne porządki") w połączeniu z wizualnym przedstawieniem grupy podlegającej ochronie (flaga społeczności osób transpłciowych), aby zachęcać osoby z tej grupy do samobójstwa. W przypadku postanowienia dotyczącego Ormian w Azerbejdżanie rada podkreśliła znaczenie kontekstu przy ustalaniu, czy wyrażenie rozważane w tej sprawie miało być skierowane do grupy o cechach chronionych. Choć post dotyczył wojny, to zagrożenia, przed którymi stoją osoby transpłciowe w Polsce, pokazują, że społeczności mogą doświadczać trudnych sytuacji i bez konfliktu wojennego. Jak podano powyżej, 1 na 5 osób transpłciowych doświadczyła przemocy fizycznej lub seksualnej w ciągu pięciu lat poprzedzających rok 2019; zgłoszeń takich było ponad dwukrotnie więcej niż w przypadku innych grup LGBTI.

43. Rada wyraża zaniepokojenie faktem, że osoby sprawdzające wstępnie treści nie zwróciły uwagi na kontekst i w rezultacie doszły do wniosku, że treść nie narusza standardów. Rada pragnie zalecić zmiany w wytycznych dotyczących egzekwowania standardów społeczności dotyczących mowy nienawiści, ale jednocześnie podkreśla, że post naruszał zasady w ówcześnie opublikowanej wersji. Obydwa stwierdzenia zawarte w poście zachęcały osoby transpłciowe do samobójstwa. Wniosek ten znajduje również potwierdzenie w opisie użytkownika, który opublikował post. Samoidentyfikacja użytkownika jako transfoba sama w sobie stanowiłaby naruszenie Poziomu 2 standardu dotyczącego mowy nienawiści (zakaz „przyznania się do braku tolerancji ze względu na chronioną cechę"). Meta musi poprawić skuteczność egzekwowania zasad w zakresie nawoływania do nienawiści wobec osób LGBTQIA+, zwłaszcza gdy posty zawierają obrazy i tekst, których interpretacja wymaga informacji kontekstowych. Jak zauważa GLAAD (PC-16027), Meta „konsekwentnie nie egzekwuje swoich zasad podczas weryfikacji zgłoszeń treści zawierających złośliwą kreatywność".

44. Ponadto rada jest zaniepokojona stwierdzeniem Mety, że nieusunięcie treści przez osoby sprawdzające treści jest „równoznaczne ze ścisłym stosowaniem wewnętrznych wytycznych".

Twierdzenie Mety wskazuje, że wewnętrzne wytyczne dla osób sprawdzających treści w niedostatecznym stopniu odzwierciedlają sposoby, w jakie połączenie tekstu i obrazu w poście w mediach społecznościowych może przedstawiać grupę zdefiniowaną na podstawie tożsamości płciowej jej członków. Rada uważa, że wytyczne mogą nie wystarczyć, aby osoby sprawdzające masowo zgłaszane treści były w stanie osiągnąć odpowiedni stopień egzekwowania zasad w przypadku treści skierowanych do grup podlegających ochronie, które są przedstawiane wizualnie, ale bez podania nazwy lub zilustrowania ich postaciami ludzkimi. Meta nie zezwoliła radzie na publikację dodatkowych informacji, które umożliwiłyby bardziej szczegółową dyskusję na temat sposobów poprawy egzekwowania zasad w przypadku tego typu treści. Rada uważa jednak, że Meta powinna zmodyfikować wytyczne tak, aby zapewnić właściwe zrozumienie wizualnych przedstawień tożsamości płciowej podczas oceny treści pod kątem ataków. Rada podkreśla, że proponując takie rozwiązania, nie ma na celu osłabienia ochrony Mety wyłącznie do koncepcji, instytucji, idei, praktyk lub przekonań. Rada nalega raczej, aby Meta wyjaśniła, że posty atakujące osoby nie muszą przedstawiać postaci ludzkich.

Samobójstwo i samookaleczenie

45. Rada uznała, że treści w omawianej sprawie naruszyły standard społeczności dotyczący samobójstw i samookaleczeń. Zasady te zabraniają publikowania „treści, które promują, zachęcają, koordynują lub zawierają instrukcje dotyczące samobójstwa lub samookaleczenia". Zgodnie z wewnętrznymi wytycznymi, które Meta przekazuje osobom sprawdzającym treści, „promowanie" definiuje się jako „pozytywne mówienie o danej kwestii". Rada zgadza się z ostatecznym wnioskiem Mety, że omawiana treść stanowi popieranie śmierci przez samobójstwo, skierowane do chronionej grupy, a zatem zachęca do samobójstwa.

46. Rada uważa również, że standard społeczności dotyczący samobójstw i samookaleczeń powinien wyraźniej zabraniać publikowania treści promujących samobójstwo lub zachęcających do samobójstwa i skierowanych do możliwej do zidentyfikowania grupy osób, a nie tylko pojedynczych osób z takiej grupy. Meta przekazała radzie, że w zasadach nie ma rozróżnienia tych dwóch form treści. Biorąc jednak pod uwagę wyzwania, przed jakimi stanęły osoby sprawdzające treści podczas klasyfikacji stwierdzenia zachęcającego do samobójstwa skierowanego do grupy osób, rada nalega, by Meta dodała informację, że zasada zabrania publikowania treści promujących lub zachęcających do samobójstwa i skierowanych do możliwej do zidentyfikowania grupy osób. Meta powinna objaśnić tę kwestię na stronie

zawierającej zasady dotyczące samobójstw i samookaleczeń, a także w powiązanych wewnętrznych wytycznych dla osób sprawdzających treści.

II. Egzekwowanie zasad

47. Rada uważa, że systemy automatycznego ustalania priorytetów sprawdzania treści wdrożone przez Metę znacząco wpłynęły na egzekwowanie zasad w tym przypadku. Dziesięć spośród 12 zgłoszeń użytkowników dotyczących postu zostało automatycznie zamkniętych przez zautomatyzowane systemy Mety. Dwa z trzech odwołań użytkowników od decyzji Mety zostały automatycznie zamknięte przez zautomatyzowane systemy Mety. Rada wyraża zaniepokojenie faktem, że historia sprawy udostępniona radzie zawiera liczne oznaki naruszenia zasad, co sugeruje, że zasady Mety nie są egzekwowane w wystarczający sposób.

48. Rada zauważyła, że wiele zgłoszeń użytkowników zostało zamkniętych ze względu na praktyki Mety dotyczące moderowania treści, polegające na łączeniu wielu zgłoszeń dotyczących tej samej treści. Pierwsze zgłoszenie użytkownika dotyczące mowy nienawiści nie zostało poddane sprawdzeniu przez człowieka ze względu na „niską wagę i wskaźnik rozpowszechniania". Kolejne zgłoszenia dotyczące mowy nienawiści nie były priorytetowo przesyłane do sprawdzenia przez człowieka, ponieważ w przypadku wielu zgłoszeń dotyczących tej samej treści Meta „usuwa powtarzające się zgłoszenia, aby zapewnić spójność decyzji osób sprawdzających treści i działań następczych". Rada przyznaje, że usuwanie powtarzających się zgłoszeń jest rozsądną praktyką w przypadku moderacji masowo zgłaszanych treści. Rada zauważa jednak, że w takiej sytuacji kładzie się nacisk na wstępne ustalenia dokonywane na podstawie zgłoszenia, ponieważ od tego zależy również los wszystkich zgłoszeń zgrupowanych z pierwotnym zgłoszeniem.

49. Rada uważa, że Meta powinna priorytetowo potraktować poprawę dokładności zautomatyzowanych systemów, które zarówno egzekwują zasady dotyczące treści, jak i przesyłają je do sprawdzenia, szczególnie w przypadku treści, które potencjalnie mogą mieć wpływ na społeczność osób LGBTQIA+. Taka poprawa zdolności zautomatyzowanych systemów do rozpoznawania zakodowanego języka i obrazów kontekstowych rozważanych w tym przypadku niewątpliwie poprawiłaby egzekwowanie zasad w przypadku treści skierowanych również przeciwko innym grupom objętym ochroną. Zdaniem rady na przykład opis użytkownika, który zawierał przyznanie się do transfobii, mógł zostać uwzględniony jako jeden z istotnych sygnałów przy ustalaniu wagi zgłoszenia w celu podjęcia decyzji, czy przesłać

priorytetowo treść do sprawdzenia, czy też podjąć działania zgodnie z zasadami. Sygnał taki może uzupełnić istniejące analizy behawioralne i analizy sieci społecznościowych, które Meta może wykorzystywać do wykrywania treści potencjalnie naruszających zasady.

50. Ponadto rada pragnie podkreślić, że dla Meta ważne byłoby zapewnienie poprawnej kalibracji zautomatyzowanych systemów oraz przeszkolenie osób sprawdzających treści pod kątem skutecznej oceny masowo zgłaszanych postów dotyczących społeczności LGBTQIA+. Rada wyraża zaniepokojenie obecnym podejściem Meta, zgodnie z którym osoby sprawdzające treści, którym powierzono ocenę odwołań, często wydają się mieć ten sam poziom wiedzy specjalistycznej, co osoby dokonujące pierwotnej oceny treści. Rada uważa, że Meta powinna zająć się bardziej intensywnie opracowaniem i rozwijaniem klasyfikatorów, które pozwalają identyfikować treści potencjalnie naruszające standardy i mające wpływ na społeczność osób LGBTQIA+, a także oznaczają je jako konieczne do sprawdzenia przez człowieka. Mowa nienawiści, zwłaszcza treści o najwyższej wadze (Poziom 1 zasad Meta), powinna zawsze być traktowana priorytetowo pod kątem sprawdzenia przez człowieka. Rada sugeruje również wzmocnienie ulepszeń procesów poprzez: i) wprowadzenie ulepszonych szkoleń dla osób sprawdzających treści na temat przemocy związanej z tożsamością płciową; ii) utworzenie grupy zadaniowej ds. doświadczeń osób transpłciowych i niebinarnych na platformach Mety; oraz iii) utworzenie wyspecjalizowanej grupy ekspertów zajmujących się weryfikacją treści związanych z kwestiami wpływającymi na społeczność osób LGBTQIA+. Chociaż fakty w omawianej sprawie odnoszą się konkretnie do przemocy doświadczanej przez osoby transpłciowe na Facebooku, rada zachęca również Metę do zbadania, w jaki sposób można poprawić egzekwowanie zasad w przypadku nienawistnych treści wpływających na inne grupy objęte ochroną.

51. Chociaż Rada wydaje jedynie dwa formalne zalecenia, to pragnie podkreślić, że wynika to ze związku problemów wynikłych w omawianej sprawie z egzekwowaniem zasad, a nie samymi zasadami. W tej sprawie rada zidentyfikowała co najmniej pięć przesłanek wskazujących na szkodliwość treści: (1) zawarte w poście odniesienia do „zasłon, które wieszają się same”; (2) odniesienie do „wiosennych porządków <3”; (3) samookreślenie użytkownika jako „transfoba” w kontekście kraju, w którym zgłaszany jest wysoki poziom wrogości wobec społeczności osób LGBTQIA+; (4) liczba zgłoszeń i odwołań dotyczących treści; oraz (5) liczba zgłoszeń i odwołań w stosunku do stopnia rozpowszechnienia treści. Rada wyraża zaniepokojenie faktem, że Meta przeoczyła te sygnały, i uważa, że sugeruje to niedostateczne egzekwowanie zasad. Zarząd pozostaje na stanowisku, że Meta powinna

zdecydowanie i kreatywnie przemyśleć, w jaki sposób zapewnić egzekwowanie zasad ochrony osób LGBTQIA+ na swoich platformach w świetle publikowanych przez spółkę zobowiązań do takiej ochrony.

8.2 Zgodność z obowiązkami Mety w zakresie przestrzegania praw człowieka

Wolność wyrażania opinii (Artykuł 19 MPPOiP [ICCPR])

52. Artykuł 19, ustęp 2 Międzynarodowego paktu praw obywatelskich i politycznych (MPPOiP) (International Covenant on Civil and Political Rights, ICCPR) stanowi, co następuje: „Każdy człowiek ma prawo do swobodnego wyrażania opinii; prawo to obejmuje swobodę poszukiwania, otrzymywania i rozpowszechniania wszelkich informacji i poglądów, bez względu na granice państwowe, ustnie, pismem lub drukiem, w postaci dzieła sztuki bądź w jakikolwiek inny sposób według własnego wyboru". Komentarz ogólny nr 34 (2011) stanowi dodatkowo, że ochrona opinii obejmuje także opinie, które można uznać za „wysoce obraźliwe" (ustęp 11).

53. Jeżeli dane państwo narzuca ograniczenia wypowiedzi, muszą one spełniać wymogi legalności, uzasadnionego celu oraz konieczności i proporcjonalności (art. 19 ustęp 3 MPPOiP). Wymogi te są często określane jako „test trzyczęściowy". Zarząd wykorzystuje te ramy do interpretacji dobrowolnych zobowiązań Meta w zakresie praw człowieka, zarówno w odniesieniu do decyzji dotyczącej pojedynczej treści poddanej weryfikacji, jak i tego, jakie ma to znaczenie dla szerszego podejścia Mety do zarządzania treścią. Jak objaśnił Specjalny Sprawozdawca ds. Promocji i Ochrony Prawa do Wolności Opinii i Wypowiedzi, chociaż „przedsiębiorstwa nie mają obowiązków takich jak rządy, ich wpływ wymaga od nich oceny tego samego rodzaju kwestii dotyczących ochrony prawa ich użytkowników do wolności wypowiedzi" (A/74/486, ustęp 41).

I. Legalność (klarowność i dostępność zasad)

54. Zasada legalności wynikająca z międzynarodowych praw człowieka wymaga, aby zasady ograniczające wypowiedzi były jasne i publicznie dostępne (Komentarz ogólny nr 34, ustęp 25). Zasady ograniczające wypowiedzi „nie mogą przyznawać nieograniczonej swobody ograniczania wolności wypowiedzi osobom odpowiedzialnym za ich egzekwowanie" i muszą „zapewniać odpowiedzialnym za ich egzekwowanie wytyczne wystarczające, by umożliwić im

ustalenie, jakie rodzaje wypowiedzi są odpowiednio ograniczone, a jakie nie" (*ibid.*). Specjalny Sprawozdawca ds. Promocji i Ochrony Prawa do Wolności Opinii i Wypowiedzi stwierdził, że w odniesieniu do zasad dotyczących wypowiedzi w Internecie wytyczne te powinny być jasne i konkretne (A/HRC/38/35, ustęp 46). Osoby korzystające z platform Meta powinny mieć dostęp do tych zasad i je rozumieć, a osoby sprawdzające treści powinny mieć jasne wytyczne dotyczące ich egzekwowania.

55. Rada stwierdza, że zakazy Meta dotyczące „mowy nienawiści lub popierania nienawiści" w formie pisemnej lub wizualnej skierowanej przeciwko grupom o cechach chronionych i wyrażeń, że chroniona cecha nie powinna istnieć oraz wypowiedzi promujących lub zachęcających do samobójstwa i samookaleczenia, są wystarczająco klarowne.

56. Rada pragnie jednak odnotować, że Meta mogłaby poprawić staranność egzekwowania zasad w odniesieniu do zasad mających zastosowanie do omawianej sprawy, zapewniając jaśniejsze wytyczne dla osób sprawdzających treści, jak omówiono w ustępie 8.1 powyżej. Meta powinna zapewnić dodatkowe wyjaśnienia dotyczące tego, że wizualne przedstawienie tożsamości płciowej, np. poprzez flagę, nie musi przedstawiać postaci ludzkich, aby stanowiło atak w świetle zasad dotyczących mowy nienawiści. Meta powinna również wyjaśnić, że namawianie całej grupy (a nie tylko pojedynczej osoby) do popełnienia samobójstwa narusza zasady dotyczące samobójstw i samookaleczeń.

II. Uzasadniony cel

57. Wszelkie ograniczenia wypowiedzi powinny służyć jednemu z uzasadnionych celów MPPOiP, które obejmują także „prawa innych osób". W kilku wcześniejszych postanowieniach rada stwierdziła, że zasady Mety dotyczące mowy nienawiści, których celem jest ochrona ludzi przed szkodami spowodowanymi mową nienawiści, mają uzasadniony cel uznawany przez międzynarodowe standardy prawa dotyczące praw człowieka (patrz np. postanowienie dotyczące rysunku miasta Knin). Ponadto Rada stwierdza, że w tym przypadku zasady dotyczące samobójstw i samookaleczeń dotyczące treści zachęcających do samobójstwa lub samookaleczenia służą uzasadnionym celom, jakim jest ochrona prawa człowieka do korzystania z najwyższego osiągalnego poziomu zdrowia fizycznego i psychicznego (ICESCR 12) i prawa do życia (art. 5 MPPOiP). W przypadkach podobnych do omawianej sprawy, w których grupa o charakterze chronionym jest zachęcana do popełnienia samobójstwa,

zasady dotyczące samobójstw i samookaleczeń chronią również prawa człowieka do równości i braku dyskryminacji (art. 2 ustęp 1 MPPOiP).

III. Konieczność i proporcjonalność

58. Zasada konieczności i proporcjonalności stanowi, że wszelkie ograniczenia wolności słowa „muszą być na tyle adekwatne, by spełniać swoją funkcję ochronną; muszą być najmniej inwazyjnym instrumentem spośród tych, które mogą pełnić funkcję ochronną; [i] muszą być proporcjonalne do chronionego interesu” (Komentarz ogólny nr 34, ustęp 34).

59. Analizując ryzyko stwarzane przez treści zawierające przemoc, rada zazwyczaj kieruje się sześciopunktowym testem opisanym w Planie Działania z Rabatu, który dotyczy szerzenia nienawiści na tle narodowym, rasowym lub religijnym, stanowiącej podżeganie do wrogości, dyskryminacji lub przemocy. Na podstawie oceny stosownych czynników, w szczególności treści i formy wypowiedzi, intencji osoby wypowiadającej się oraz kontekstu opisanego szerzej poniżej, rada stwierdza, że usunięcie treści jest zgodne z obowiązkami Mety w zakresie praw człowieka, ponieważ treść ta stwarza nieuchronne i prawdopodobne zagrożenie. Usunięcie treści stanowi konieczne i proporcjonalne ograniczenie wypowiedzi w celu ochrony prawa do życia i prawa do korzystania z najwyższego osiągalnego poziomu zdrowia fizycznego i psychicznego szerszej społeczności osób LGBTQIA+, a w szczególności osób transpłciowych w Polsce.

60. Choć rada już wcześniej podkreśliła znaczenie przywrócenia postów zawierających obraźliwe określenia wobec osób ze społeczności LGBTQIA+ w celu przeciwdziałania dezinformacji (patrz postanowienie w sprawie przywrócenia postu ze słowami w języku arabskim), nie ma to zastosowania w tym przypadku. Post nie zawiera również wypowiedzi politycznych ani wartych opublikowania (patrz postanowienie w sprawie protestów w Kolumbii). W omawianym w tej sprawie poście zamieszczono się zdjęcie flagi społeczności osób transpłciowych powieszonej jako zasłony z opisem, że „zasłony wieszają się same”. Zdaniem ekspertów, z którymi konsultowała się rada, używanie zasłon – zarówno w formie wizualnej, jak i tekstowej – nie wydaje się powtarzającym się zakodowanym językiem skierowanym przeciwko społeczności osób transpłciowych. Niemniej jednak, jak wspomniano powyżej, zjawisko „złośliwej kreatywności”, czyli używania języka i strategii przedstawiania koncepcji w nowatorski sposób w celu wyrażania nienawiści i nękania, zaczęło być cechą charakterystyczną trendów w treściach skierowanych przeciwko osobom transpłciowym. Rada uznała, że treści w omawianej sprawie wpisują się w ten trend. Chociaż w poście wykorzystano obrazy, które niektórzy uznali za „humorystyczne” (o czym świadczą emotikony „Haha”), nadal można go interpretować jako brutalne i prowokacyjne stwierdzenie skierowane przeciwko społeczności osób transpłciowych. Humor i satyrę można oczywiście

wykorzystać do przesuwania ustalonych granic uzasadnionej krytyki, ale nie mogą one stanowić swego rodzaju przykrywki dla mowy nienawiści. W poście poruszany jest temat wysokiego wskaźnika samobójstw wśród społeczności osób transpłciowych wyłącznie w celu promowania tego faktu.

61. Rozważając intencje twórcy treści, rada zauważyła, że w opisie konta stwierdza on otwarcie, że jest „transfobem". Chociaż Meta dopiero później rozważyła konsekwencje tego stwierdzenia dla samej treści sprawy, rada uznaje je za wysoce istotne dla określenia intencji użytkownika. Stanowiłoby ono również podstawę do usunięcia treści jako naruszenie zasad dotyczących mowy nienawiści (Poziom 2). W poście opisano także akt samobójczej śmierci osób transpłciowych jako „wiosenne porządki", dodając obok opisu emotikon serca. W świetle takiego poparcia samobójstwa skierowanego do całej grupy rada stwierdza, że zamiarem postu było zachęcanie do dyskryminacji i przemocy, co wynika zarówno z treści postu, jak i użytego zdjęcia i towarzyszącego mu tekstu oraz podpisu. Treść postu w omawianej sprawie nie tylko zachęca osoby transpłciowe do podejmowania wobec siebie brutalnych działań, ale także nawołuje innych do dyskryminacji i wrogości wobec osób transpłciowych. Potwierdza to fakt, że najczęściej stosowanym emotikonem reakcji wśród innych użytkowników wchodzących w interakcję z tą treścią był emotikon śmiechu, „Haha".

62. Rada pragnie również zwrócić uwagę na znaczące ryzyko offline, przed którym stoi polska społeczność osób LGBTQIA+, w postaci nasilających się ataków w drodze działań legislacyjnych i administracyjnych, a także retoryki politycznej ze strony osobistości z instytucji rządowych i wpływowych osób publicznych. Od 2020 roku Polska konsekwentnie osiąga najniższe wyniki w zakresie praw osób LGBTQIA+ wśród państw członkowskich UE w badaniach prowadzonych przez ILGA-Europe. Należy również zauważyć, że w Polsce nie przewidziano przepisów dotyczących mowy nienawiści i przestępstw z nienawiści w odniesieniu do społeczności osób LGBTQIA+, do czego wzywają Polskę między innymi takie organizacje, jak ILGA-Europe i Amnesty International. Co więcej, narastająca retoryka skierowana przeciwko społeczności osób LGBTQIA+ w języku polskim na Facebooku, sygnalizowana przez zewnętrznych ekspertów i w licznych komentarzach publicznych, nie jest zjawiskiem odosobnionym. Wiele organizacji i instytucji wyraziło zaniepokojenie powszechnością wypowiedzi skierowanych przeciwko społeczności LGBTQIA+ w mediach społecznościowych. Niezależny ekspert ONZ ds. orientacji seksualnej i tożsamości płciowej (IE SOGI), Victor Madrigal-Borloz, stwierdził, że poziom przemocy i dyskryminacji wobec osób nienormatywnych płciowo i osób transpłciowych „obraża ludzkie sumienie". W badaniach i raportach GLAAD stwierdzono, że „osobom ze społeczności LGBTQ wyrządza się w Internecie realne szkody, w tym… powodujące traumę psychologiczną wynikającą z ciągłego narażenia na obelgi i nienawistne zachowanie". Treści takie jak omawiany post, zwłaszcza publikowane na dużą skalę, mogą przyczynić się do powstania środowiska, w którym pogłębiają się i tak wszechobecne szkody wynikające z samobójstw w społeczności osób transpłciowych.

Ponadto treści normalizujące agresywne wypowiedzi skierowane przeciwko osobom transpłciowym, jak ma to miejsce w omawianym poście, mogą przyczynić się zarówno do trwającego kryzysu zdrowia psychicznego, który wpływa na społeczność osób transpłciowych, jak i do wzrostu przemocy wymierzonej w tę społeczność offline.

9. Postanowienie rady Oversight Board

63. Rada Oversight Board uchyla pierwotną decyzję Mety o pozostawieniu postu.

10. Zalecenia

A. Zasady dotyczące treści

1. Na stronie standardów dotyczących samobójstw i samookaleczeń Meta powinna dodać informację, że zasada zabrania publikowania treści promujących samobójstwo lub zachęcających do samobójstwa i skierowanych do możliwej do zidentyfikowania grupy osób.

Rada uzna to zalecenie za wdrożone w momencie, gdy dostępna publicznie treść standardu społeczności dotyczącego samobójstw i samookaleczeń będzie odzwierciedlać proponowaną zmianę.

B. Egzekwowanie

2. Wewnętrzne wytyczne Mety przekazywane osobom sprawdzającym masowo zgłaszane treści powinny zostać zmodyfikowane tak, aby upewnić się, że wizualne przedstawienia dotyczące tożsamości płciowej, które nie zawierają postaci ludzkiej, ale obejmują flagi, były rozumiane jako przedstawienie grupy określonej na podstawie tożsamości płciowej jej członków. Taka modyfikacja umożliwiłaby dodatkowe objaśnienie instrukcji dotyczących egzekwowania zasad w przypadku tej formy zgłaszanej masowo treści, gdy zawiera ona atak naruszający zasady.

Rada uzna to zalecenie za wdrożone w momencie, gdy Meta przedstawi radzie zmiany w swoich wewnętrznych wytycznych.

Informacje dotyczące procedur:

Postanowienia rady Oversight Board są przygotowywane przez panele składające się z pięciu członków i zatwierdzane głosami większości rady. Postanowienia rady niekoniecznie odzwierciedlają osobiste poglądy wszystkich jej członków.

Na potrzeby tego postanowienia w imieniu rady zlecono niezależne badanie. Radę wspomagał niezależny instytut badawczy z siedzibą na Uniwersytecie w Göteborgu, w skład którego wchodzi zespół ponad 50 socjologów z sześciu kontynentów, a także ponad 3200 ekspertów z całego świata. Analizę dostarczyła także Memetica, organizacja zajmująca się badaniami open source nad trendami w mediach społecznościowych. Ekspertyzę językową zapewniła firma Lionbridge Technologies, LLC, której specjaliści władają biegle ponad 350 językami i pracują w 5000 miejscowości na całym świecie.