Snap Inc.

Nowojorski raport dotyczący Regulaminu świadczenia usług

1 lipca – 30 września 2025 r.

Przesłano: 1 stycznia 2026 r.

Nowojorski raport dotyczący Regulaminu świadczenia usług (1 lipca – 30 września 2025 r.)
Snap Inc.

Snap Inc. przekazuje niniejszy raport dotyczący Regulaminu świadczenia usług do Prokuratora Generalnego Nowego Jorku, zgodnie z ogólnym prawem gospodarczym Nowego Jorku (New York General Business Law), art. 42, §§ 1100–1104. Niniejszy nowojorski raport dotyczący Regulaminu świadczenia usług obejmuje okres od 1 lipca 2025 r. do 30 września 2025 r. (III kwartał 2025 r.).

Nasze Warunki (NY Gen. Bus. L. §§ 1101, 1102)

Dokładamy wszelkich starań, aby Snapchat był miejscem bezpiecznym i pełnym rozrywki, stawiającym na kreatywność i ekspresję. Wszyscy snapchaterzy muszą przestrzegać naszego Regulaminu świadczenia usług, w tym naszych Wytycznych dla społeczności (w niniejszym raporcie łącznie określanych jako „Warunki"). Dodatkowe informacje na temat tego, jak moderujemy treści i egzekwujemy nasze zasady można znaleźć w naszej serii objaśnień do Wytycznych dla społeczności. Zawierają one opis naszej polityki dotyczącej moderowania, egzekwowania zasad i odwołań, a także dodatkowe informacje dotyczące każdej kategorii treści i zachowań zabronionych w naszych Wytycznych dla społeczności.

Oprócz naszych Warunków w naszym Centrum bezpieczeństwa udostępniamy informacje i zasoby związane z bezpieczeństwem, w tym wskazówki dotyczące tego, jak zgłaszać domniemane naruszenia naszych Warunków lub inne wątpliwości dotyczące bezpieczeństwa w naszej usłudze.

Wszystkie dokumenty wymienione powyżej są załączone do niniejszego raportu w języku angielskim oraz są dostępne na naszej stronie internetowej w dwunastu najbardziej powszechnych językach innych niż angielski. Są to języki, którymi posługują się osoby znające angielski w ograniczonym stopniu w Nowym Jorku, gdzie oferujemy usługę Snapchata.

Zasady i praktyki moderowania treści (NY Gen. Bus L. § 1102)

Nasze Warunki zabraniają kategorii treści, do których odnosi się **NY Gen. Bus. L. § 1102(c)**, wymienionych poniżej:

Kategoria treści	**Odpowiadająca kategoria treści zabronionych przez nasze Wytyczne dla społeczności**	**Odpowiednie definicje i zasady, zgodnie z naszą serią objaśnień do Glosariusza terminologii stosowanej w raporcie przejrzystości oraz Wytycznych dla społeczności**
Mowa nienawiści lub rasizm	Mowa nienawiści (która podlega pod treści nienawistne, terroryzm i brutalny ekstremizm)	Mowa nienawiści lub treści poniżające, zniesławiające albo promujące dyskryminację czy przemoc ze względu na rasę, kolor skóry, kastę, pochodzenie etniczne, narodowość, wyznanie, orientację seksualną, tożsamość płciową, niepełnosprawność, a także status weterana, status społeczno-gospodarczy, wiek, wagę lub ciążę. Zasady te zabraniają na przykład używania obelg rasowych, etnicznych, mizoginicznych lub homofobicznych. Zabraniają również publikowania memów ośmieszających lub nawołujących do dyskryminacji grupy chronionej, a także wszelkiego rodzaju nadużyć przyjmujących postać celowego stosowania nieaktualnych danych osobowych lub form płciowych wobec osób, które zmieniły płeć. Mowa nienawiści obejmuje również popieranie sprawców — lub lekceważenie ofiar — ludzkich tragedii (takich jak ludobójstwo, apartheid czy niewolnictwo). Inne zabronione treści nienawistne zawierają wykorzystanie symboli nienawiści, czyli wszelkich obrazów, których intencją jest przedstawienie nienawiści lub dyskryminacji wobec innych osób. Więcej informacji znajdziesz w naszym objaśnieniu dotyczącym treści nienawistnych, terroryzmu i brutalnego ekstremizmu.
Ekstremizm lub radykalizacja	Terroryzm i brutalny ekstremizm (które podlegają pod treści nienawistne, terroryzm i brutalny ekstremizm).	Treści promujące terroryzm lub inne brutalne i przestępcze działania osób lub grup w celu realizacji celów ideologicznych. Zasady te zakazują również wszelkich treści, które promują lub wspierają zagraniczne organizacje terrorystyczne albo ekstremistyczne grupy szerzące nienawiść (wskazane przez wiarygodnych, zewnętrznych ekspertów), a także rekrutację do takich organizacji, lub namawianie do uczestnictwa

		w brutalnych działaniach ekstremistycznych. Więcej informacji znajdziesz w naszym objaśnieniu dotyczącym treści nienawistnych, terroryzmu i brutalnego ekstremizmu.
Dezinformacja lub informacje wprowadzające w błąd	Fałszywe informacje (które podlegają pod praktyki szkodliwe, fałszywe lub wprowadzające w błąd).	Obejmują fałszywe lub wprowadzające w błąd treści, które wyrządzają szkody lub mają złe intencje. Są to między innymi zaprzeczanie zaistnieniu tragicznych wydarzeń, bezpodstawne twierdzenia medyczne, podważanie integralności procesów obywatelskich lub manipulowanie treściami w celach fałszowania lub wprowadzania w błąd (w tym za pomocą generatywnej sztucznej inteligencji lub poprzez wprowadzającą w błąd edycję). Aby uzyskać więcej informacji, zapoznaj się z naszym objaśnieniem dotyczącym szkodliwych, fałszywych lub wprowadzających w błąd praktyk.
Napastowanie	(1) napastowanie i nękanie oraz (2) napastowanie seksualne (które podpada pod treści o charakterze seksualnym) (w poniższej tabeli łącznie określane jako „Napastowanie”).	Odnoszą się do wszelkich niepożądanych zachowań, które mogłyby spowodować u zwykłej osoby stres emocjonalny. Są to między innymi przemoc słowna, groźby lub wszelkie zachowania mające na celu zawstydzenie, zakłopotanie lub upokorzenie innej osoby. Więcej informacji znajdziesz w naszym objaśnieniu dotyczącym napastowania i nękania. Ponadto Wytyczne dla społeczności dotyczące treści o charakterze seksualnym zabraniają wszelkich form napastowania seksualnego. Mogą one obejmować składanie innym użytkownikom niewłaściwych propozycji, udostępnianie im drastycznych i niechcianych treści lub wysyłanie nieprzyzwoitych próśb lub zaproszeń seksualnych. Więcej informacji znajdziesz w naszym objaśnieniu dotyczącym treści o charakterze seksualnym.
Zagraniczna ingerencja polityczna	Fałszywe informacje (które podlegają pod szkodliwe, fałszywe lub wprowadzające w błąd praktyki i obejmują m.in. źródła zagraniczne lub polityczne).	Nasza definicja fałszywych informacji znajduje się powyżej. Do tej kategorii należy również podawanie się za inną osobę, mające miejsce wtedy, gdy konto fałszywie udaje powiązanie z inną osobą lub marką. Aby uzyskać więcej informacji, zapoznaj się z naszym objaśnieniem dotyczącym szkodliwych, fałszywych lub wprowadzających w błąd praktyk.

Nasze objaśnienie dotyczące moderowania, egzekwowania zasad i odwołań oraz objaśnienie dotyczące znaczących szkód zawierają szczegółowe informacje na następujące tematy:

- w jaki sposób moderujemy treści za pomocą zautomatyzowanych narzędzi i weryfikacji przez człowieka,
- jak reagujemy na zgłoszenia użytkowników dotyczące domniemanych naruszeń naszych Wytycznych dla społeczności, oraz
- jak podejmujemy działania egzekucyjne wobec poszczególnych elementów treści i użytkowników naruszających nasze Wytyczne dla społeczności.

Informacje dotyczące naruszania naszych Warunków (1 lipca – 30 września 2025 r.) (NY Gen. Bus. L. § 1102)

Poniżej przedstawiamy szczegółowe informacje na temat naruszeń naszych Wytycznych dla społeczności, które zostały nam zgłoszone w aplikacji lub zostały automatycznie wykryte przez nasze systemy w okresie od 1 lipca do 30 września 2025 r., zgodnie z ogólnym prawem gospodarczym Nowego Jorku (NY Gen. Bus. L. § 1102). Podane dane mają charakter ogólnoświatowy. Poniższe dane nie obejmują zgłoszeń dokonywanych poza aplikacją Snapchata (to znaczy za pośrednictwem strony wsparcia i poczty e-mail), które stanowią mniej niż 1% całkowitej liczby zgłoszeń.

O ile nie określono inaczej, terminy użyte w niniejszej sekcji są definiowane zgodnie z naszym Glosariuszem terminologii dotyczącej przejrzystości.

Kategoria naruszenia	Sposób oznaczenia	Łączna liczba oznaczonych treści lub kont[1]	Treści, wobec których podjęto działania egzekucyjne[2] po weryfikacji przez człowieka	Treści, wobec których podjęto działania egzekucyjne z wykorzystaniem zautomatyzowanych narzędzi	Unikalne konta, wobec których podjęto działania egzekucyjne[3] po weryfikacji przez człowieka	Unikalne konta, wobec których podjęto działania egzekucyjne z wykorzystaniem zautomatyzowanych narzędzi	Odwołania od decyzji o zablokowaniu konta po weryfikacji przez człowieka	Odwołania od decyzji o zablokowaniu konta wymuszonym przez zautomatyzowane narzędzia	Konta przywrócone w następstwie odwołania[4] (początkowo zablokowane po weryfikacji przez człowieka)	Konta przywrócone w następstwie odwołania (początkowo zablokowane przez zautomatyzowane narzędzia)	Wskaźnik wyświetleń treści naruszających zasady (VVR)[5], wobec których podjęto działania egzekucyjne po weryfikacji przez człowieka	VVR dla treści, wobec których podjęto działania egzekucyjne z wykorzystaniem zautomatyzowanych narzędzi	Wskaźnik unikalnych widzów wyświetlających treści naruszające zasady[6], wobec których podjęto działania egzekucyjne po weryfikacji przez człowieka	Wskaźnik unikalnych widzów wyświetlających treści naruszające zasady, wobec których podjęto działania egzekucyjne z wykorzystaniem zautomatyzowanych narzędzi
Mowa nienawiści	Zgłoszone przez człowieka	482 240	144 389	20 679	123 835	18 343	94	0	5	0	0,000205%	0,000004%	0,41%	0,01%
	Wykryte proaktywnie	3073	2441	32	2029	29	7	1164	1	15	0,000002%	0,000000%	0,00%	0,00%
Terroryzm i brutalny ekstremizm	Zgłoszone przez człowieka	199 245	559	54	468	54	199	0	1	0	0,000037%	0,000000%	0,07%	0,00%
	Wykryte proaktywnie	12 240	5901	0	3754	0	11	38	1	6	0,000000%	0,000000%	0,00%	0,00%
Fałszywe informacje	Zgłoszone przez człowieka	386 210	145	593	113	593	1	0	0	0	0,000003%	0,000000%	0,01%	0,00%
	Wykryte proaktywnie	262	10	0	10	0	1	0	0	0	0,000000%	0,000000%	0,00%	0,00%
Napastowanie	Zgłoszone przez człowieka	3 184 734	756 736	604 798	624 142	491 972	700	0	17	0	0,001135%	0,000168%	1,93%	0,35%
	Wykryte proaktywnie	12 293	8972	1583	7670	1428	17	13 442	1	97	0,000007%	0,000000%	0,02%	0,00%

(1) Całkowita liczba treści lub kont oznaczonych ze względu na potencjalne naruszenia naszych Wytycznych dla społeczności, w tym zgłoszonych do nas oraz wykrytych w ramach naszych procesów wykrywania proaktywnego. Aby podzielić te dane na kategorie treści naruszających zasady, wykorzystaliśmy ostateczny powód podjęcia działań egzekucyjnych. W przypadku gdy treści lub konto zostały oznaczone, ale nie podjęto działań egzekucyjnych, przypisujemy wskaźniki do tej kategorii domniemanego naruszenia przepisów, ze względu na którą dana treść lub konto zostały oznaczone.

(2) Liczba treści (np. Snapów, Stories), wobec których podjęto działania egzekucyjne na Snapchacie. „Działanie egzekucyjne” odnosi się do działania podjętego wobec danej treści lub konta (np. usunięcie, ostrzeżenie, zablokowanie).

(3) Liczba unikalnych kont, wobec których podjęto działania egzekucyjne na Snapchacie. Na przykład jeśli wobec jednego konta podjęto wielokrotne działania egzekucyjne z różnych powodów (np. użytkownik otrzymał ostrzeżenie z powodu zamieszczania fałszywych informacji, a następnie jego konto zostało zablokowane za nękanie innego użytkownika), konto to zostanie policzone w tym wskaźniku jako jedno. Jak powyżej, „działanie egzekucyjne” odnosi się do działania podjętego wobec danej treści lub konta (np. usunięcie, ostrzeżenie, zablokowanie).

(4) Przywracamy wyłącznie te konta, które zostaną uznane przez naszych moderatorów za niesłusznie zablokowane.

(5) Wskaźnik wyświetleń treści naruszających zasady to odsetek wyświetleń Stories i Snapów zawierających treści naruszające zasady, w proporcji do wszystkich wyświetleń Stories i Snapów na Snapchacie. (Snap to zdjęcie lub wideo wykonane za pomocą kamery Snapchata. Więcej informacji znajduje się tutaj). Na przykład, jeśli nasza wartość VVR wynosi 0,03%, oznacza to, że na każde 10 000 wyświetleń Snapów i

Stories na Snapchacie, 3 zawierały treści naruszające nasze zasady. Wskaźnik ten umożliwia nam zrozumienie, jaki odsetek wyświetleń Snapów i Stories na Snapchacie dotyczy treści, które naruszają nasze Wytyczne dla społeczności (zgłoszonych lub przeciwko którym podjęte zostały działania prewencyjne).

(6) Wskaźnik unikalnych widzów wyświetlających treści naruszające zasady to odsetek unikalnych widzów, którzy widzieli naruszające zasady Stories i/lub Snapy, w proporcji do wszystkich unikalnych aktywnych użytkowników w okresie objętym raportem (tj. w III kwartale 2025 r.). Na przykład, jeśli nasz wskaźnik unikalnych widzów wyświetlających treści naruszające zasady wynosi 0,03%, oznacza to, że na każde 10 000 użytkowników aktywnych w danym okresie na Snapchacie, 3 z nich widziało Stories i/lub Snapy naruszające nasze zasady. Wskaźnik ten umożliwia nam zrozumienie, jaki odsetek użytkowników na Snapchacie natrafia na Stories i/lub Snapy naruszające nasze Wytyczne dla społeczności (zgłoszone lub przeciwko którym podjęte zostały działania prewencyjne).

Dodatkowe informacje

Aby uzyskać dodatkowe informacje na temat podejścia firmy Snap do bezpieczeństwa, prywatności i przejrzystości, odwiedź nasze Centrum prywatności, bezpieczeństwa i zasad oraz stronę o raportach przejrzystości.