Snap Inc.
Informe de Condiciones de Servicio de California

1 de julio al 30 de septiembre de 2023

Fecha de nueva presentación: 7 de mayo de 2024

Informe de Condiciones de servicio de California (1 de julio al 30 de septiembre de 2023) (Nueva presentación)
Snap Inc.

Motivo de la nueva presentación

De conformidad con la sección 22677 del Código de Negocios y Profesiones de California, Snap Inc. ("Snap") presenta este Informe de Condiciones de Servicio ante el Fiscal General de California. Esta es una nueva presentación del primer Informe de Condiciones de Servicio de California de Snap, que abarca el período comprendido entre el 1 de julio de 2023 y el 30 de septiembre de 2023 (tercer trimestre de 2023), con la intención de aclarar dos omisiones involuntarias. En primer lugar, el informe se actualiza para reflejar que durante el período de informe relevante, Snap tenía políticas que prohibían la interferencia política extranjera como parte de sus Pautas para la comunidad. En segundo lugar, el informe se actualiza para incluir la Explotación sexual infantil como una categoría separada y distinta de infracción. Este cambio da lugar a actualizaciones de ciertos datos, que también se reflejan en la nueva presentación. El Informe de Condiciones de Servicio de Snap del cuarto trimestre de 2023, que se envió el 1 de abril de 2024, ya refleja esta categoría adicional de Explotación sexual infantil.

Nuestras Condiciones (Código de Negocios y Profesiones de California, §§22677(a)(1) y (4)(E))

Nos esforzamos por proporcionar un entorno seguro y divertido para la creatividad y la expresión en Snapchat. Todos los usuarios de Snapchat deben cumplir nuestras Condiciones de servicio, incluidas nuestras Pautas para la comunidad (en conjunto, las "**Condiciones**").

Hay contexto adicional sobre cómo moderamos el contenido y hacemos cumplir nuestras políticas disponible en nuestra Serie de documentos explicativos sobre las Pautas para la comunidad, que incluye una descripción de nuestras políticas de Moderación, penalización y apelaciones e información adicional sobre todas las categorías de contenido que prohíben nuestras Pautas para la comunidad.

En nuestro Centro de seguridad, también proporcionamos información y recursos relacionados con la seguridad, lo que incluye orientación sobre cómo denunciar infracciones a nuestras Condiciones u otros problemas de seguridad relacionados con nuestro servicio.

Dichos documentos se adjuntan a este informe en inglés y están disponibles en nuestro sitio web en todos los idiomas de umbral de Medi-Cal en los que ofrecemos Snapchat.

Políticas y prácticas de moderación de contenido (Código de Negocios y Profesiones de California, §§22677(a)(3)-(4))

Nuestras Condiciones prohíben las categorías de contenido mencionadas en la Sección 22677(a)(3), según se describe a continuación:

Categoría de contenido mencionada en la Sección 22677(a)	**Categoría de contenido correspondiente prohibida por nuestras Pautas para la comunidad**	**Definiciones y políticas relevantes, según lo descrito en nuestro Glosario del informe de transparencia y la serie de documentos explicativos sobre las Pautas para la comunidad**
Incitación al odio o racismo	Incitación al odio (según se describe en Contenido de odio, terrorismo y extremismo violento)	Contenido que denigra o promueve la discriminación contra una persona o grupo de personas basándose en su raza, color, casta, etnia, origen nacional, religión, orientación sexual, identidad de género, discapacidad, condición de veterano, condición de inmigrante, estado socioeconómico, edad, peso o embarazo. Para obtener más información, revisá nuestro Documento explicativo sobre contenido de odio, terrorismo y extremismo violento.
Extremismo o radicalización	Terrorismo y extremismo violento (según se describe en Contenido de odio, terrorismo y extremismo violento)	Contenido que promueva o apoye el terrorismo u otros actos delictivos violentos cometidos por personas o grupos de personas para fomentar objetivos ideológicos, como los de naturaleza política, religiosa, social, racial o ambiental. Incluye cualquier contenido que promueva o apoye cualquier organización terrorista extranjera o grupo de odio extremista violento, así como contenido que promueva el reclutamiento para dichas organizaciones o actividades extremistas violentas. Para obtener más información, revisá nuestro Documento explicativo sobre contenido de odio, terrorismo y extremismo violento.

Desinformación o información errónea	Información falsa (según se describe en Información falsa o engañosa dañina)	Incluye contenido falso o engañoso que causa daño o es malintencionado, como negar la existencia de eventos trágicos, declaraciones médicas sin fundamento, socavar la integridad de los procesos cívicos o manipular contenido con fines falsos o engañosos. Para obtener más información, revisá nuestro Documento explicativo sobre información falsa o engañosa dañina.
Hostigamiento	Acoso y hostigamiento	Se refiere a cualquier comportamiento no deseado que podría causar que una persona común experimente angustia emocional, como abuso verbal, acoso sexual o atención sexual no deseada. Esta categoría también incluye compartir o recibir imágenes íntimas no consensuadas (NCII). Para obtener más información, revisá nuestro Documento explicativo sobre acoso y hostigamiento.
Interferencia política extranjera	Información falsa (según se describe en Información falsa o engañosa dañina).	Consultá nuestra definición de información falsa en la sección anterior. La suplantación de identidad ocurre cuando una cuenta finge falsamente estar asociada con otra persona o marca. Para obtener más información, revisá nuestro Documento explicativo sobre información falsa o engañosa dañina.
Distribución de sustancias controladas	Drogas (según se describe en Actividades ilegales o reguladas)	Se refiere a la distribución y el uso de drogas ilegales (incluidas las píldoras falsificadas) y otras actividades ilícitas que involucren drogas. Para obtener más información, revisá nuestro Documento explicativo sobre actividades ilegales o reguladas.

En nuestro Documento explicativo sobre Moderación, penalización y apelaciones y nuestro Documento explicativo sobre daños graves se proporciona información detallada sobre los siguientes temas, entre otros:

- cómo moderamos el contenido a través de herramientas automatizadas y la revisión humana,
- cómo respondemos a las denuncias de usuarios de presuntas infracciones a nuestras Pautas para la comunidad y
- cómo penalizamos las piezas individuales de contenido y usuarios que infringen nuestras Pautas para la comunidad.

Información sobre infracciones a nuestras Condiciones (1 de julio al 30 de septiembre de 2023) (Código de Negocios y Profesiones de California, §22677(a)(5))

A continuación, proporcionamos información detallada sobre las denuncias de infracciones a nuestras Pautas para la comunidad que recibimos o que fueron detectadas automáticamente por nuestros sistemas en el período del 1 de julio al 30 de septiembre de 2023, conforme a la Sección 22677(a). Primero proporcionamos cifras globales, seguidas de cifras de Estados Unidos. Estas cifras no solo se relacionan con las categorías de contenido que infringe las pautas a las que se hace referencia en la Sección 22677(a)(3), sino, en un sentido más general, con las infracciones a las que se hace referencia en nuestras Pautas para la comunidad.[1]

Excepto cuando se especifique lo contrario, los términos utilizados en esta sección se definen conforme a nuestro Glosario de transparencia.

[1] En este informe, discriminamos los datos en: (i) categorías de contenido infractor, (ii) cómo se detectó el contenido o la cuenta (es decir, mediante una denuncia o mediante nuestras herramientas de detección automatizada) y (iii) cómo se penalizó el contenido o la cuenta (es decir, por revisores humanos o mediante herramientas automatizadas). No podemos discriminar los datos por tipo de contenido (por ejemplo, publicaciones, comentarios, mensajes, perfiles de usuario) ni por tipo de medio (por ejemplo, texto, imagen y video) en este momento, debido a que en el tercer trimestre de 2023 no monitoreábamos dichos datos a nivel global ni en los Estados Unidos, en una forma que nos permitiese extraer los datos a los fines de elaboración de un informe.

Cifras globales

Categoría de infracción	Forma en que se detectó	Contenido total o cuentas detectadas[1]	Contenido penalizado[2] por revisores humanos	Contenido penalizado mediante herramientas automatizadas	Cuentas únicas penalizadas[3] por revisores humanos	Cuentas únicas penalizadas mediante herramientas automatizadas	Apelaciones contra bloqueos de cuentas[4] penalizadas por revisores humanos	Apelaciones contra bloqueos de cuentas penalizadas mediante herramientas automatizadas	Cuentas restablecidas después de la apelación[5] (inicialmente bloqueadas por revisores humanos)	Cuentas restablecidas después de la apelación (inicialmente bloqueadas mediante herramientas automatizadas)	Tasa de visualización de contenido infractor (VVR)[6] para contenido penalizado por revisores humanos	VVR para contenido penalizado mediante herramientas automatizadas	Tasa de espectadores infractores únicos[7] para contenido penalizado por revisores humanos	Tasa de espectadores infractores únicos para contenido penalizado mediante herramientas automatizadas
Incitación al odio	Denuncia humana	189 981	45 028	257	39 567	183	206	5	11	0	0,000193 %	0,000001 %	0,44 %	0,002 %
	Detección automática	148	148	0	132	0	0	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
Terrorismo y extremismo violento	Denuncia humana	41 399	835	24	751	21	17	0	1	0	0,000005 %	0,000000 %	0,01 %	0,000 %
	Detección automática	11	11	0	11	0	0	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
Información falsa	Denuncia humana	216 219	460	10	445	9	3	0	0	0	0,000005 %	0,000000 %	0,01 %	0,000 %
	Detección automática	16	16	0	16	0	0	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
Suplantación de identidad	Denuncia humana	213 879	8040	36	8002	33	769	0	51	0	0,000002 %	0,000000 %	0,01 %	0,000 %
	Detección automática	5	5	0	5	0	0	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
	Denuncia humana	4 531 005	505 999	20 239	414 702	11 285	14 546	943	410	13	0,001143 %	0,000044 %	1,52 %	0,051 %

Acoso y hostigamiento	**Detección automática**	2523	2481	42	2268	12	78	3	7	0	0,000002 %	0,000000 %	0,00 %	0,000 %
Drogas	**Denuncia humana**	177 028	115 835	5010	84 731	4118	8331	1056	231	5	0,000536 %	0,000031 %	0,75 %	0,062 %
	Detección automática	636 008	286 538	158 894	242 067	128 763	73 446	20 420	1992	103	0,000101 %	0,000010 %	0,23 %	0,028 %
Amenazas y violencia	**Denuncia humana**	401 227	44 172	5210	34 555	3648	747	4	35	0	0,000678 %	0,000035 %	1,08 %	0,064 %
	Detección automática	410	323	11	292	6	42	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
Suicidio y autolesiones	**Denuncia humana**	85 339	15 896	56	14 637	33	18	1	5	0	0,000007 %	0,000000 %	0,01 %	0,000 %
	Detección automática	260	252	0	242	0	2	0	0	0	0,000000 %	0,000000 %	0,00 %	0,000 %
Correo no deseado	**Denuncia humana**	1 254 516	311 954	514 111	269 775	312 043	7287	128	108	1	0,000858 %	0,000106 %	1,28 %	0,218 %
	Detección automática	50 890	15 636	35 254	14 084	21 633	443	96	7	0	0,000004 %	0,000029 %	0,01 %	0,021 %
Armas	**Denuncia humana**	48 967	6129	568	4831	409	214	45	6	1	0,000035 %	0,000001 %	0,06 %	0,002 %
	Detección automática	123 755	40 106	66 208	32 953	51 275	612	995	25	8	0,000022 %	0,000006 %	0,06 %	0,016 %
	Denuncia humana	228 900	68 618	4582	52 689	2351	3989	508	111	4	0,000526 %	0,000018 %	0,87 %	0,029 %

Otros productos regulados	**Detección automática**	9967	9925	42	8668	21	389	25	27	1	0,000010 %	0,000000 %	0,03 %	0,001 %
Contenido sexual	**Denuncia humana**	2 146 825	794 265	398 293	580 110	249 112	60 534	4233	747	19	0,004442 %	0,001858 %	3,08 %	1,392 %
	Detección automática	397 538	150 421	194 379	98 190	111 567	11 177	1392	125	10	0,000061 %	0,000011 %	0,10 %	0,019 %
Explotación sexual infantil	**Denuncia humana**	389 163	113 454	2547	96 106	1949	13 677	68	2059	11	0,000300 %	0,000020 %	0,45 %	0,017 %
	Detección automática	168 527	78 427	60 312	54 058	44 284	9124	9170	745	2015	0,000002 %	0,000000 %	0,00 %	0,001 %
Totales		11 314 506	2 614 974	1 466 085	1 920 608	910 767	205 651	39 092	6703	2191	0,008932 %	0,002172 %	5,99 %	1,694 %

Cifras de Estados Unidos

Categoría de infracción	Forma en que se detectó	Contenido total o cuentas detectadas[1]	Contenido penalizado[2] por revisores humanos	Contenido penalizado mediante herramientas automatizadas	Cuentas únicas penalizadas[3] por revisores humanos	Cuentas únicas penalizadas mediante herramientas automatizadas	Apelaciones contra bloqueos de cuentas[4] penalizadas por revisores humanos	Apelaciones contra bloqueos de cuentas penalizadas mediante herramientas automatizadas	Cuentas restablecidas después de la apelación[5] (inicialmente bloqueadas por revisores humanos)	Cuentas restablecidas después de la apelación (inicialmente bloqueadas mediante herramientas automatizadas)	Tasa de visualización de contenido infractor (VVR)[6] para contenido penalizado por revisores humanos	Tasa de visualización de contenido infractor (VVR) para contenido penalizado mediante herramientas	Tasa de espectadores infractores únicos[7] para contenido penalizado por revisores humanos	Tasa de espectadores infractores únicos para contenido penalizado mediante herramientas automatizadas
Incitación al odio	**Denuncia humana**	74 256	26 254	184	22 888	127	118	0	7	0	0,0004208 %	0,0000048 %	1,316 %	0,015 %
	Detección automática	86	86	0	79	0	0	0	0	0	0,0000003 %	0.0000000 %	0,001 %	0,000 %
	Denuncia humana	10 901	197	6	190	4	4	0	0	0	0,0000062 %	0,0000001 %	0,020 %	0,000 %

Terrorismo y extremismo violento	**Detección automática**	6	6	0	6	0	0	0	0	0	0.0000000 %	0.0000000 %	0,000 %	0,000 %
	Denuncia humana	47 421	235	3	223	3	0	0	0	0	0,0000072 %	0,0000002 %	0,023 %	0,001 %
Información falsa	**Detección automática**	10	10	0	10	0	0	0	0	0	0.0000000 %	0.0000000 %	0,000 %	0,000 %
	Denuncia humana	54 948	2461	13	2442	11	241	0	16	0	0,0000001 %	0.0000000 %	0,000 %	0,000 %
Suplantación de identidad	**Detección automática**	2	2	0	2	0	0	0	0	0	0.0000000 %	0.0000000 %	0,000 %	0,000 %
	Denuncia humana	1 134 660	166 787	4658	140 939	3385	3987	89	173	9	0,0017937 %	0,0000227 %	4,261 %	0,051 %
Acoso y hostigamiento	**Detección automática**	1189	1186	3	1092	2	28	1	4	0	0,0000043 %	0.0000000 %	0,014 %	0,000 %
	Denuncia humana	76 888	54 098	1922	39 227	1655	3439	166	96	0	0,0014146 %	0,0000292 %	2,821 %	0,111 %
Drogas	**Detección automática**	369 835	170 066	117 427	142 887	93 734	37 681	11 458	909	58	0,0002741 %	0,0000334 %	0,980 %	0,141 %
	Denuncia humana	117 412	16 571	1432	13 448	1057	316	0	22	0	0,0006518 %	0,0000524 %	1,691 %	0,134 %
Amenazas y violencia	**Detección automática**	222	167	10	153	5	26	0	0	0	0,0000007 %	0,0000001 %	0,002 %	0,000 %
	Denuncia humana	29 226	8027	8	7583	8	6	0	3	0	0,0000126 %	0.0000000 %	0,040 %	0,000 %

Suicidio y autolesiones	**Detección automática**	159	153	0	146	0	0	0	0	0	0.0000000 %	0.0000000 %	0,000 %	0,000 %
Correo no deseado	**Denuncia humana**	580 657	137 514	360 649	124 895	223 162	1997	19	22	0	0,0009065 %	0,0000995 %	2,147 %	0,326 %
	Detección automática	15 974	6304	9670	6126	6276	122	1	2	0	0,0000037 %	0,0000129 %	0,016 %	0,030 %
Armas	**Denuncia humana**	17 212	1742	80	1604	72	66	9	3	0	0,0000382 %	0,0000009 %	0,142 %	0,004 %
	Detección automática	99 084	32 206	56 158	26 788	43 961	449	209	17	4	0,0000886 %	0,0000241 %	0,345 %	0,101 %
Otros productos regulados	**Denuncia humana**	73 261	13 629	306	11 770	210	340	20	23	1	0,0004930 %	0,0000038 %	1,482 %	0,012 %
	Detección automática	3539	3534	5	3173	3	63	0	10	0	0,0000098 %	0.0000000 %	0,041 %	0,000 %
Contenido sexual	**Denuncia humana**	584 728	221 552	127 380	163 531	84 979	16 924	824	233	8	0,0057545 %	0,0025898 %	8,864 %	4,121 %
	Detección automática	109 214	39 790	44 859	27 328	29 015	3926	238	38	5	0,0001110 %	0,0000145 %	0,327 %	0,042 %
Explotación sexual infantil	**Denuncia humana**	109 155	25 071	245	22 045	181	13 677	68	2059	11	0,0001473 %	0,0000049 %	0,290 %	0,011 %
	Detección automática	33 376	12 754	11 707	9686	8503	9124	9170	745	2015	0,0000009 %	0,0000001 %	0,002 %	0,000 %
Totales		3 543 421	940 402	736 725	725 906	486 592	205 651	39 092	6703	2191	0,0121398 %	0,0028934 %	16,290 %	4,772 %

(1) Número total de piezas de contenido o cuentas que se marcaron por posibles infracciones a nuestras Pautas para la comunidad, incluidas las que nos denunciaron y las que se detectaron mediante nuestras herramientas automatizadas. Para desglosar estos datos en categorías de contenido infractor, usamos el motivo definitivo de penalización en los casos en que se tomó una acción de penalización. En los casos en que se denunció el contenido o la cuenta pero no se tomaron medidas de penalización, atribuimos las métricas a la categoría de presunta infracción por la que se marcó el contenido o la cuenta.
(2) El número de piezas de contenido (por ejemplo, Snaps, Historias) que se penalizaron en Snapchat. "Penalización" se refiere a una medida que se toma contra una pieza de contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).
(3) El número de cuentas únicas que se penalizaron en Snapchat. Por ejemplo, si una sola cuenta fue sancionada varias veces por varias razones (por ejemplo, se advirtió a un usuario por publicar información falsa y luego se eliminó su cuenta por acosar a otro usuario), solo una cuenta se calcularía en esta métrica. Como se mencionó anteriormente, "penalización" se refiere a una medida que se toma contra una pieza de contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).
(4) Los usuarios solo pueden presentar apelaciones contra un bloqueo de cuenta.
(5) Solo restablecemos las cuentas que nuestros moderadores determinan que fueron bloqueadas incorrectamente.
(6) La tasa de visualización de contenido infractor (VVR) es el porcentaje de vistas de Historias y Snaps que contenían contenido infractor, como una proporción de todas las vistas de Historias y Snaps en Snapchat. Por ejemplo, si nuestra VVR es del 0,03 %, eso significa que por cada 10 000 visualizaciones de Snaps e Historias en Snapchat, 3 tenían contenido que infringió nuestras políticas. Esta métrica nos permite comprender qué porcentaje de vistas en Snapchat provienen de contenido que infringe nuestras Pautas para la comunidad (que se denunció o penalizó de manera proactiva).
(7) La tasa de espectadores infractores únicos es el porcentaje de espectadores únicos que vieron contenido infractor, como una proporción de usuarios únicos activos durante el período del informe, es decir, el tercer trimestre de 2023. Por ejemplo, si nuestra tasa de espectadores únicos infractores es del 0,03 %, eso significa que, por cada 10 000 usuarios activos durante el período correspondiente en Snapchat, 3 espectadores vieron contenido que infringió nuestras políticas. Esta métrica nos permite comprender qué porcentaje de usuarios de Snapchat se topan con contenido que infringe nuestras Pautas para la comunidad (que se denunció o penalizó de manera proactiva).

Información adicional

Aunque no se requiere conforme a la Sección 22677, también creemos que es valioso proporcionar nuestros tiempos de respuesta promedio (TAT) para responder a las denuncias y apelaciones. Definimos TAT como el tiempo entre el momento en que nuestros equipos de Confianza y Seguridad o las Herramientas automatizadas reciben una denuncia por primera vez (generalmente cuando se envía una denuncia o se detecta a través de medios automatizados) y la marca de tiempo de la última medida de penalización. Si ocurren varias rondas de revisión, el tiempo final se calcula en la última medida que se tomó. Teniendo esto en cuenta, nuestro TAT promedio global para denuncias de contenido y cuentas es de aproximadamente 6 minutos.

Para obtener información adicional sobre el enfoque de Snap sobre la seguridad, la privacidad y la transparencia, visitá nuestro Centro de privacidad y seguridad y nuestra página Acerca de la transarencia en la presentacion de informes.