Snap Inc.
Informe de Términos del servicio de California

Del 1 de julio al 30 de septiembre de 2023

Se presentó nuevamente: 07 de mayo de 2024

PROYECTO – A/C PRIVILEGIADO Y CONFIDENCIAL

Informe de Términos del servicio de California (1 de julio al 30 de septiembre de 2023) (nueva presentación)
Snap Inc.

Razón para hacer una nueva presentación

De conformidad con la Sección 22677 del Código de Empresas y Profesiones de California, Snap Inc. («Snap») presenta este Informe de Términos del servicio al Fiscal General de California. Esta es una nueva presentación del primer Informe de Términos del 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. Primero, este informe se actualiza para reflejar que, durante el período abarcado por el informe, Snap tenía políticas que prohibían la interferencia política extranjera como parte de sus Pautas para la comunidad. En segundo lugar, este informe se actualiza para incluir la explotación sexual infantil como una categoría de incumplimiento separada y distinta. Este cambio resulta en actualizaciones de ciertos datos, lo que también se refleja en esta nueva presentación. El Informe de los Términos del servicio del cuarto trimestre de 2023 de Snap, que se presentó el 1 de abril de 2024, ya refleja esta categoría adicional de explotación sexual infantil.

Nuestros Términos (Cal. Bus. & Prof. Code, §§22677(a)(1) y (4)(E))

Nos esforzamos por ofrecer un entorno seguro y divertido para la creatividad y la expresión en Snapchat. Todos los usuarios de Snapchat deben cumplir nuestros Términos del servicio, incluidas nuestras Pautas para la comunidad (conjuntamente, las «**Condiciones**»).

En nuestra serie explicativa sobre las Pautas para la comunidad, se encuentra más información sobre cómo moderamos el contenido y hacemos cumplir nuestras políticas, que incluye una descripción de nuestras políticas de moderación, aplicación y apelaciones, así como información adicional sobre cada categoría de contenido y conducta prohibida por nuestras Pautas para la comunidad.

También proporcionamos información y recursos relacionados con la seguridad en nuestro Centro de seguridad, incluida la orientación sobre cómo denunciar infracciones a nuestras Condiciones u otras preocupaciones de seguridad sobre nuestro servicio.

Estos documentos están adjuntos a este informe en inglés y están disponibles en nuestro sitio web en todos los idiomas que alcanzan el umbral de Medi-Cal y en los que ofrecemos Snapchat.

Políticas y prácticas de moderación de contenido (Cal. Bus. & Prof. Code, §§22677(a)(3)-(4))

Nuestras Condiciones prohíben las categorías de contenido mencionadas en la Sección 22677(a)(3), tal y como se detalla 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 previsto en nuestro Glosario del informe de transparencia y la serie explicativa sobre las Pautas para la comunidad**
Discurso de odio o racismo	Discurso de odio (que entra en el contenido de odio, terrorismo y extremismo violento)	Contenido que denigra o fomenta la discriminación contra una persona o grupo de personas por su raza, color, casta, etnia, origen nacional, religión, orientación sexual, identidad de género, discapacidad, estado de veterano, estado de inmigración, estado socioeconómico, edad, peso o estado de embarazo. Para más información, consulta nuestro documento explicativo sobre Contenido que incita al odio, terrorismo y extremismo violento.
Extremismo o radicalización	Terrorismo y extremismo violento (que entra en el contenido que incita al odio, el terrorismo y el extremismo violento)	Contenido que promueva o apoye el terrorismo u otros actos delictivos violentos cometidos por individuos o grupos para promover objetivos ideológicos, como los de naturaleza política, religiosa, social, racial o medioambiental. 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 tales organizaciones o actividades extremistas violentas. Para más información, consulta nuestro documento explicativo sobre Contenido que incita al odio, terrorismo y extremismo violento.

Desinformación o información errónea	Información falsa (que entra en la categoría de Información falsa perjudicial o engañosa)	Incluye contenido falso o engañoso que causa daño o es malicioso, como negar la existencia de eventos trágicos, afirmaciones médicas sin fundamento, socavar la integridad de los procesos cívicos, manipular contenido con fines falsos o engañosos. Para más información, consulta nuestro documento explicativo sobre Información falsa o engañosa nociva.
Acoso	Acoso y acoso virtual	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 el intercambio o la recepción de imágenes íntimas no consensuadas (NCII, por sus siglas en inglés). Para más información, consulta nuestro documento explicativo sobre Acoso y bullying.
Injerencia política extranjera	Información falsa (que entra en la categoría Información falsa o engañosa perjudicial).	Para obtener nuestra definición de información falsa, consulta la sección anterior. La suplantación de identidad se produce cuando una cuenta finge falsamente estar asociada con otra persona o marca. Para más información, consulta nuestro documento explicativo sobre Información falsa o engañosa nociva.
Distribución controlada de sustancias	Drogas (que entran en las actividades ilegales o reguladas)	Se refiere a la distribución y el uso de drogas ilegales (incluidas las píldoras falsificadas) y otras actividades ilícitas relacionadas con las drogas. Para más información, consulta nuestro documento explicativo sobre Actividades ilegales o reguladas.

Nuestro explicativo sobre moderación, aplicación y apelaciones y el explicativo sobre daños graves proporcionan información detallada sobre estos temas, entre otros:

- cómo moderamos el contenido a través de herramientas automatizadas y revisión humana,
- cómo respondemos a las denuncias de los usuarios sobre supuestas infracciones de nuestras Pautas para la comunidad, y
- cómo tomamos medidas contra las piezas individuales de contenido y los usuarios que infringen nuestras Pautas para la comunidad.

Información sobre infracciones de nuestras Condiciones (1 de julio al 30 de septiembre de 2023) (Cal. Bus. & Prof. Code, §22677(a)(5))

A continuación, proporcionamos información detallada sobre las infracciones de nuestras Pautas para la comunidad que se denunciaron o que nuestros sistemas detectaron automáticamente en el período del 1 de julio al 30 de septiembre de 2023, de conformidad con la Sección 22677(a). Primero proporcionamos cifras globales, seguidas de cifras de los Estados Unidos. Estas cifras se relacionan no solo con las categorías de contenido infractor mencionadas en la Sección 22677(a)(3), sino en términos más generales con las infracciones a las que se hace referencia en nuestras Pautas para la comunidad.[1]

Excepto que se especifique lo contrario, los términos utilizados en esta sección se definen de acuerdo con nuestro Glosario de transparencia.

[1] En este informe, hemos desagregado los datos en: (i) categorías de contenido infractor, (ii) cómo se denunció el contenido o la cuenta (es decir, mediante un informe o nuestras herramientas de detección automatizada), y (iii) cómo se tomaron medidas de cumplimiento sobre el contenido o la cuenta (es decir, por revisores humanos o mediante herramientas automatizadas). No podemos desglosar los datos por tipo de contenido (por ejemplo, publicaciones, comentarios, mensajes, perfiles de usuario) o por tipo de medio (por ejemplo, texto, imagen, vídeo) en este momento, porque no estábamos realizando un seguimiento de estos datos a nivel mundial ni en Estados Unidos a partir del tercer trimestre de 2023, de una manera que nos permitiera extraer estos datos a efectos de la presentación de informes.

Cifras globales

Categoría de incumplimiento	Forma de detección	Contenido total o cuentas detectados[1]	Contenido con medidas aplicadas[2] por revisores humanos	Contenido con medidas aplicadas por herramientas automatizadas	Cuentas únicas con medidas aplicadas[3] por revisores humanos	Cuentas únicas con medidas aplicadas por herramientas automatizadas	Apelaciones contra bloqueos de cuenta[4] aplicados por revisores humanos	Apelaciones contra bloqueos de cuenta aplicados por herramientas automatizadas	Cuentas restablecidas tras una apelación[5] (bloqueadas inicialmente por revisores humanos)	Cuentas restablecidas tras una apelación (inicialmente bloqueadas por herramientas automatizadas)	Tasa de visualizaciones no permitidas (VVR)[6] para el contenido con medidas aplicadas por revisores humanos	VVR para contenido con medidas aplicadas por herramientas automatizadas	Tasa única de espectadores infractores[7] para el contenido con medidas aplicadas por revisores humanos	Tasa única de espectadores infractores para el contenido con medidas aplicados por herramientas automatizadas
Discurso de odio	Informe humano	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	0,00000 %	0,00000 %	0,00 %	0,000 %
Terrorismo y extremismo violento	Informe humano	41 399	835	24	751	21	17	0	1	0	0,000005 %	0,00000 %	0,01 %	0,000 %
	Detección automática	11	11	0	11	0	0	0	0	0	0,00000 %	0,00000 %	0,00 %	0,000 %
Información falsa	Informe humano	216 219	460	10	445	9	3	0	0	0	0,000005 %	0,00000 %	0,01 %	0,000 %
	Detección automática	16	16	0	16	0	0	0	0	0	0,00000 %	0,00000 %	0,00 %	0,000 %
Suplantación de identidad	Informe humano	213 879	8040	36	8002	33	769	0	51	0	0,000002 %	0,00000 %	0,01 %	0,000 %
	Detección automática	5	5	0	5	0	0	0	0	0	0,00000 %	0,00000 %	0,00 %	0,000 %
Acoso y acoso virtual	Informe humano	4 531 005	505 999	20 239	414 702	11 285	14 546	943	410	13	0,001143 %	0,000044 %	1,52 %	0,051 %
	Detección automática	2523	2481	42	2268	12	78	3	7	0	0,000002 %	0,00000 %	0,00 %	0,000 %
	Informe humano	177 028	115 835	5010	84 731	4118	8331	1056	231	5	0,000536 %	0,000031 %	0,75 %	0,062 %

Drogas	**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	**Informe humano**	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	0,00000 %	0,00000 %	0,00 %	0,000 %
Autolesiones y suicidio	**Informe humano**	85 339	15 896	56	14 637	33	18	1	5	0	0,000007 %	0,00000 %	0,01 %	0,000 %
	Detección automática	260	252	0	242	0	2	0	0	0	0,00000 %	0,00000 %	0,00 %	0,000 %
Spam	**Informe humano**	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	**Informe humano**	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	0,00022 %	0,000006 %	0,06 %	0,016 %
Otros bienes regulados	**Informe humano**	228 900	68 618	4582	52 689	2351	3989	508	111	4	0,000526 %	0,000018 %	0,87 %	0,029 %
	Detección automática	9967	9925	42	8668	21	389	25	27	1	0,000010 %	0,00000 %	0,03 %	0,001 %
Contenido sexual	**Informe humano**	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 %
	Informe humano	389 163	113 454	2547	96 106	1949	13 677	68	2059	11	0,000300 %	0,000020 %	0,45 %	0,017 %

Explotación sexual infantil	**Detección automática**	168 527	78 427	60 312	54 058	44 284	9124	9170	745	2015	0,000002 %	0,00000 %	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 incumplimiento	**Forma de detección**	**Contenido total o cuentas detectados[1]**	**Contenido con medidas aplicadas[2] por revisores humanos**	**Contenido con medidas aplicadas por herramientas automatizadas**	**Cuentas únicas con medidas aplicadas[3] por revisores humanos**	**Cuentas únicas con medidas aplicadas por herramientas automatizadas**	**Apelaciones contra bloqueos de cuenta[4] aplicados por revisores humanos**	**Apelaciones contra bloqueos de cuenta aplicados por herramientas automatizadas**	**Cuentas restablecidas tras una apelación[5] (bloqueadas inicialmente por revisores humanos)**	**Cuentas restablecidas tras una apelación (inicialmente bloqueadas por herramientas automatizadas)**	**Tasa de visualizaciones no permitidas (VVR)[6] para el contenido con medidas aplicadas por revisores humanos**	**Tasa de visualizaciones no permitidas (VVR) para el contenido con medidas aplicadas por**	**Tasa única de espectadores infractores[7] para el contenido con medidas aplicadas por revisores humanos**	**Tasa única de espectadores infractores para el contenido con medidas aplicados por herramientas automatizadas**
Discurso de odio	**Informe humano**	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 %
Terrorismo y extremismo violento	**Informe humano**	10 901	197	6	190	4	4	0	0	0	0,0000062 %	0,0000001 %	0,020 %	0,000 %
	Detección automática	6	6	0	6	0	0	0	0	0	0,0000000%	0,0000000 %	0,000 %	0,000 %
Información falsa	**Informe humano**	47 421	235	3	223	3	0	0	0	0	0,0000072 %	0,0000002 %	0,023 %	0,001 %
	Detección automática	10	10	0	10	0	0	0	0	0	0,0000000%	0,0000000 %	0,000 %	0,000 %
	Informe humano	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 %
Acoso y acoso virtual	**Informe humano**	1 134 660	166 787	4658	140 939	3385	3987	89	173	9	0,0017937 %	0,0000227 %	4,261 %	0,051 %
	Detección automática	1189	1186	3	1092	2	28	1	4	0	0,0000043 %	0,0000000 %	0,014 %	0,000 %
Drogas	**Informe humano**	76 888	54 098	1922	39 227	1655	3439	166	96	0	0,0014146 %	0,0000292 %	2,821 %	0,111 %
	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 %
Amenazas y violencia	**Informe humano**	117 412	16 571	1432	13 448	1057	316	0	22	0	0,0006518 %	0,0000524 %	1,691 %	0,134 %
	Detección automática	222	167	10	153	5	26	0	0	0	0,0000007 %	0,0000001 %	0,002 %	0,000 %
Autolesiones y suicidio	**Informe humano**	29 226	8027	8	7583	8	6	0	3	0	0,0000126 %	0,0000000 %	0,040 %	0,000 %
	Detección automática	159	153	0	146	0	0	0	0	0	0,0000000%	0,0000000 %	0,000 %	0,000 %
Spam	**Informe humano**	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 %	0,000129 %	0,016 %	0,030 %
Armas	**Informe humano**	17 212	1742	80	1 604	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 %
	Informe humano	73 261	13 629	306	11 770	210	340	20	23	1	0,0004930 %	0,0000038 %	1,482 %	0,012 %

Otros bienes regulados	**Detección automática**	3539	3534	5	3173	3	63	0	10	0	0,0000098 %	0,0000000 %	0,041 %	0,000 %
Contenido sexual	**Informe humano**	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	**Informe humano**	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 contenidos o cuentas que se señalaron por posibles infracciones de nuestras Pautas para la comunidad, incluidas las que denunciaron a nosotros y las detectadas a través de nuestras herramientas automatizadas. Para desglosar estos datos en categorías de contenido infractor, hemos utilizado el motivo principal por el cual se tomó una medida de cumplimiento. Cuando se señaló el contenido o la cuenta, pero no se tomaron acciones para aplicar medidas de cumplimiento, atribuimos las métricas a la categoría de presunto incumplimiento para la que se señaló el contenido o la cuenta.

(2) El número de piezas de contenido (por ejemplo, snaps, historias) sobre los que se aplicaron medidas en Snapchat. La «aplicación de medidas» se refiere a una acción tomada contra un contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).

(3) El número de cuentas únicas contra las que se aplicaron las medidas en Snapchat. Por ejemplo, si se aplicaron medidas contra una misma cuenta varias veces por diversos motivos (por ejemplo, se hizo una advertencia a un usuario por publicar información falsa y más adelante se eliminó su cuenta por haber acosado a otro usuario), solo se computaría una cuenta en esta métrica. Como se ha mencionado anteriormente, la «aplicación de medidas» se refiere a una acción tomada contra un contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).

(4) Los usuarios solo pueden enviar apelaciones contra un bloqueo de cuenta.

(5) Solo restablecemos cuentas que nuestros moderadores determinen que estaban bloqueadas incorrectamente.

(6) La tasa de visualizaciones no permitidas es el porcentaje de visualizaciones de historias y snaps con contenido infractor, como proporción del total de visualizaciones 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 incluían contenidos que infringían nuestras políticas. Esta métrica nos permite comprender qué porcentaje de visualizaciones en Snapchat proceden de contenidos que infringen nuestras Pautas para la comunidad (sobre los que hubo una denuncia o se aplicaron medidas de forma proactiva).

(7) La tasa de espectadores únicos infractores es el porcentaje de espectadores únicos que vieron contenido infractor, como proporción de los usuarios únicos activos durante todo el período sobre el que se informa, 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 relevante en Snapchat, 3 espectadores vieron contenidos que infringen nuestras políticas. Esta métrica nos permite comprender qué porcentaje de usuarios de Snapchat se encuentran con contenidos que infringen nuestras Pautas para la comunidad (los cuales se denunciaron o se les aplicaron medidas de forma proactiva).

Información adicional

Aunque no lo exige la Sección 22677, también creemos que es valioso proporcionar nuestros tiempos de respuesta medios (TAT, por sus siglas en inglés) para responder a las denuncias y apelaciones. Definimos TAT como el momento en que nuestros equipos de Confianza y seguridad o las Herramientas automatizadas reciben por primera vez una denuncia (normalmente cuando una denuncia se envía o se detecta a través de medios automatizados) hasta la última marca de tiempo de la acción de cumplimiento. Si se producen varias rondas de revisión, se computa como tiempo final el de la última acción tomada. Con eso en mente, nuestro TAT medio global para las denuncias de contenido y cuentas es de aproximadamente 6 minutos.

Para obtener información adicional sobre el enfoque de Snap con respecto a la seguridad, la privacidad y la transparencia, visita nuestro Centro de privacidad y seguridad y nuestra página Acerca de la presentación de informes de transparencia.