Snap Inc.:
Informe de las Condiciones de servicio de California

1 de julio - 30 de septiembre de 2023

Reenviado: 7 de mayo de 2024

Informe de las Condiciones de servicio de California (1 de julio - 30 de septiembre de 2023) (reenvío)
Snap Inc.

Motivo del reenvío

De conformidad con la Sección 22677 del Código de Empresas y Profesiones de California, Snap Inc. ("Snap") presenta el presente Informe de Condiciones de servicio ante el Fiscal General de California. Este es un reenvío 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), destinado a aclarar dos omisiones inadvertidas. En primer lugar, este 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, este informe se actualiza para incluir la explotación sexual infantil como una categoría de incumplimiento separada y distinta. Este cambio se traduce en actualizaciones de ciertos datos, que también se reflejan en este reenvío. El informe de Condiciones de servicio del cuarto trimestre de 2023 de Snap, que se envió el 1 de abril de 2024, ya refleja esta categoría adicional de explotación sexual infantil.

Nuestras Condiciones (Cód. de Emp. y Prof. de Cal., §§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**").

Encontrarás más información sobre cómo moderamos el contenido y aplicamos nuestras políticas en nuestra Serie de documentos explicativos de las Pautas para la comunidad, que incluye una descripción de nuestras políticas de moderación, penalización y apelaciones, e información adicional sobre cada categoría de contenido prohibido por nuestras Pautas para la comunidad.

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

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

Políticas y prácticas de moderación de contenidos (Cód. de Emp. y Prof. de Cal., §§22677(a)(3)-(4))

Nuestras Condiciones prohíben las categorías de contenido a las que se hace referencia en la sección 22677(a)(3), de la siguiente manera:

Categoría de contenido a la que se hace referencia en la sección 22677(a)	**Categoría correspondiente de contenido prohibida por nuestras Pautas para la comunidad**	**Definiciones y políticas relevantes, según lo proporcionado en nuestro Glosario de informes de transparencia y la serie de explicaciones de las Pautas para la comunidad**
Discurso de odio o racismo	Discurso de odio (que se incluye dentro de contenido de odio, terrorismo y extremismo violento)	Contenido que denigre o promueva la discriminación hacia una persona o grupo de personas por su raza, color, casta, etnia, origen nacional, religión, orientación sexual, identidad de género, discapacidad, condición de veterano, estado de inmigración, estado socioeconómico, edad, peso o estado de embarazo. Para más información, consulta nuestra explicación sobre el contenido que incita al odio, el terrorismo y el extremismo violento.
Extremismo o radicalización	Terrorismo y extremismo violento (que cae bajo el contenido de odio, terrorismo y extremismo violento)	Contenido que promueva o apoye el terrorismo u otros actos delictivos violentos cometidos por individuos y/o grupos para promover objetivos ideológicos, como los de naturaleza política, religiosa, social, racial o ambiental. Incluye cualquier contenido que promueva o apoye a cualquier organización terrorista extranjera o grupo de odio extremista violento, así como el contenido que promueva el reclutamiento para dichas organizaciones o actividades extremistas violentas. Para más información, consulta nuestra explicación sobre el contenido que incita al odio, el terrorismo y el extremismo violento.

Desinformación o información errónea	Información falsa (que cae dentro de la categoría de información falsa o engañosa perjudicial)	Incluye contenido falso o engañoso que cause daños o sea malicioso, como negar la existencia de eventos trágicos, afirmaciones médicas sin fundamento, socavar la integridad de los procesos cívicos o manipular contenido con fines falsos o engañosos. Para más información, consulte nuestra explicación sobre la información falsa o engañosa que sea perjudicial.
Acoso	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 el intercambio o la recepción de imágenes íntimas no consensuadas (NCII). Para más información, consulta nuestra explicación sobre acoso e intimidación.
Interferencia política extranjera	Información falsa (que se incluye en la categoría de información falsa o engañosa dañina).	Para ver nuestra definición de información falsa, consulta lo que se indica anteriormente. La suplantación de identidad se produce cuando una cuenta finge estar asociada con otra persona o marca. Para más información, consulte nuestra explicación sobre la información falsa o engañosa que sea perjudicial.
Distribución de sustancias controladas	Drogas (que se incluyen en Actividades ilegales o reguladas)	Se refiere a la distribución y uso de drogas ilegales (incluidas las píldoras falsificadas) y otras actividades ilícitas que involucran drogas. Para más información, consulta nuestra explicación sobre las actividades ilegales o reguladas.

Nuestros documentos explicativos de moderación, cumplimiento y apelaciones y explicativos de daños graves proporcionan información detallada, entre otros temas:

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

Información sobre infracciones de nuestras Condiciones (del 1 de julio al 30 de septiembre de 2023) (Cód. de Emp. y Prof. de Cal., §22677(a)(5))

A continuación proporcionamos información detallada sobre las infracciones de nuestras Pautas para la comunidad que nos fueron denunciadas o detectadas automáticamente por nuestros sistemas en el período del 1 de julio al 30 de septiembre de 2023, en cumplimiento de la Sección 22677(a). Primero proporcionamos cifras globales, seguidas de las cifras de los EE. UU. Estas cifras están relacionadas no solo con las categorías de contenido infractor a las que se hace referencia en la sección 22677(a)(3), sino de forma más general 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 desglosado los datos en: (i) categorías de contenido que infringe las normas, (ii) cómo se marcó 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 desglosar los datos por tipo de contenido (por ejemplo, publicaciones, comentarios, mensajes, perfiles de usuario) o por tipo de medio (por ejemplo, texto, imagen, video) en este momento, porque no estábamos haciendo un seguimiento de estos datos a nivel mundial o en los Estados Unidos al tercer trimestre de 2023, de una manera que nos permitiera extraerlos para fines de informes.

Cifras globales

Categoría de incumplimiento	Manera señalada	Total de contenido o cuentas marcadas[1]	Contenido penalizado[2] por revisores humanos	Contenido penalizado por herramientas automatizadas	Cuentas únicas penalizadas[3] por revisores humanos	Cuentas únicas penalizadas mediante herramientas automatizadas	Apelaciones contra bloqueos de cuentas[4] aplicados por revisores humanos	Apelaciones contra bloqueos de cuentas aplicados por herramientas automatizadas	Cuentas restablecidas tras una apelación[5] (inicialmente bloqueadas por revisores humanos)	Cuentas restablecidas tras una apelación (inicialmente bloqueadas por herramientas automatizadas)	Tasa de visualización infractor (VVR)[6] para contenido penalizado por revisores humanos	VVR para contenido penalizado por herramientas automatizadas	Tasa única de espectadores infractores[7] para contenido penalizado por revisores humanos	Tasa única de espectadores infractores para contenido penalizado 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	0,000000 %	0,000000 %	0,00 %	0,000 %
Terrorismo y extremismo violento	Informe humano	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	Informe humano	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	Informe humano	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 %
Acoso y hostigamiento	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 %	0,000000 %	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	0,000000 %	0,000000 %	0,00 %	0,000 %
Autolesiones y suicidio	**Informe humano**	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 %
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	0,000022 %	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 %	0,000000 %	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 %	0,000000 %	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 los EE. UU.

Categoría de incumplimiento	**Manera señalada**	**Total de contenido o cuentas marcadas[1]**	**Contenido penalizado[2] por revisores humanos**	**Contenido penalizado por herramientas automatizadas**	**Cuentas únicas penalizadas[3] por revisores humanos**	**Cuentas únicas penalizadas mediante herramientas automatizadas**	**Apelaciones contra bloqueos de cuentas[4] aplicados por revisores humanos**	**Apelaciones contra bloqueos de cuentas aplicados por herramientas automatizadas**	**Cuentas restablecidas tras una apelación[5] (inicialmente bloqueadas por revisores humanos)**	**Cuentas restablecidas tras una apelación (inicialmente bloqueadas por herramientas automatizadas)**	**Tasa de visualización infractor (VVR)[6] para contenido penalizado por revisores humanos**	**Tasa de visualización de contenido infractor (VVR) para contenido penalizado mediante herramientas**	**Tasa única de espectadores infractores[7] para contenido penalizado por revisores humanos**	**Tasa única de espectadores infractores para contenido penalizado por herramientas automatizadas**
	Informe humano	74 256	26 254	184	22 888	127	118	0	7	0	0,0004208 %	0,0000048 %	1,316 %	0,015 %
Discurso de odio	**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 hostigamiento	**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 %	0,0000129 %	0,016 %	0,030 %
Armas	**Informe humano**	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 %
	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 fueron denunciados por posibles infracciones de nuestras Pautas para la comunidad, incluidos los que se nos denunciaron y los que se detectaron a través de nuestras herramientas automatizadas. Para desglosar estos datos en categorías de contenido que infringe las normas, hemos utilizado el motivo último de penalización cuando se tomó una acción coercitiva. Cuando el contenido o la cuenta fue denunciado, pero no se tomaron medidas coercitivas, atribuimos las métricas a la categoría de presunto incumplimiento por la que se denunció el contenido o la cuenta.
(2) La cantidad de contenido (por ejemplo, Snaps, Historias) que se ha penalizado en Snapchat. "Penalización" se refiere a una acción tomada contra un contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).
(3) La cantidad de cuentas únicas que fueron objeto de medidas en Snapchat. Por ejemplo, si se aplicara una sola cuenta varias veces por diversos motivos (por ejemplo, si se advirtiera a un usuario por publicar información falsa y luego se eliminara más tarde por acosar a otro usuario), solo se calcularía una cuenta en esta métrica. Al igual que anteriormente, "penalización" se refiere a una acción tomada contra un contenido o una cuenta (por ejemplo, eliminación, advertencia, bloqueo).
(4) Los usuarios solo pueden presentar apelaciones contra un bloqueo de cuenta.
(5) Solo restablecimos las cuentas que nuestros moderadores determinaron que habían sido bloqueadas incorrectamente.
(6) La tasa de visualizaciones infractoras es el porcentaje de visualizaciones de Historias y Snaps que incluían contenido infractor, como proporción de todas las 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 contenían contenido que infringía nuestras políticas. Esta métrica nos permite comprender qué porcentaje de las visitas en Snapchat provienen de contenido que infringe nuestras Pautas para la comunidad (que se informó o se aplicó de forma proactiva).
(7) La tasa de espectadores únicos que infringieron las normas es el porcentaje de espectadores únicos que vieron contenido que infringió las normas, como porcentaje de usuarios únicos activos durante todo el período de informe, es decir, el tercer trimestre de 2023. Por ejemplo, si nuestra tasa de espectadores únicos que infringieron nuestras políticas es del 0,03 %, eso significa que, por cada 10 000 usuarios activos durante el período relevante en Snapchat, 3 espectadores vieron contenido que infringió nuestras políticas. Esta métrica nos permite comprender qué porcentaje de usuarios de Snapchat se encuentran con contenido que infringe nuestras Pautas para la comunidad (que se informó o se aplicó de forma proactiva).

Información adicional

Aunque no lo exige la sección 22677, también creemos que es valioso proporcionar nuestros tiempos medios de respuesta (TAT) para responder a denuncias y apelaciones. Definimos TAT como el momento que transcurre 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 o detecta una denuncia a través de medios automatizados) y la marca de tiempo de la última acción de penalización. Si se producen varias rondas de revisión, la hora final se calcula en función de la última acción realizada. Con esto en mente, nuestra media global de TAT para denuncias de contenido y cuentas es de aproximadamente 6 minutos.

Para obtener información adicional sobre el enfoque de Snap en materia de seguridad, privacidad y transparencia, visita nuestro Centro de privacidad y seguridad y nuestra página de informes de transparencia.