Snap Inc.
California Terms of Service Report

July 1 - September 30, 2023

Resubmitted: May 07, 2024

California Terms of Service Report (July 1 - September 30, 2023) (Resubmission)
Snap Inc.

Reason for resubmission

Pursuant to Section 22677 of the California Business & Professions Code, Snap Inc. ("Snap") hereby submits this Terms of Service Report to the California Attorney General. This is a resubmission of Snap's first California Terms of Service Report, covering the period between July 1, 2023, and September 30, 2023 (Q3 2023), intended to clarify two inadvertent omissions. First, this report is updated to reflect that during the relevant reporting period, Snap had in place policies prohibiting foreign political interference as part of its Community Guidelines. Second, this report is updated to include Child Sexual Exploitation as a separate and distinct category of violation. This change results in updates to certain data, which are also reflected in this resubmission. Snap's Q4 2023 Terms of Service Report, which was submitted on April 1, 2024, already reflects this additional category of Child Sexual Exploitation.

Our Terms (Cal. Bus. & Prof. Code, §§22677(a)(1) and (4)(E))

We strive to provide a safe, fun environment for creativity and expression on Snapchat. All Snapchat users must abide by our Terms of Service, including our Community Guidelines (together, "**Terms**").

Additional context about how we moderate content and enforce our policies is available in our Community Guidelines Explainer Series, which includes a description of our Moderation, Enforcement and Appeals policies and additional information regarding each category of content prohibited by our Community Guidelines.

We also provide safety-related information and resources in our Safety Center, including guidance on how to report violations of our Terms or other safety concerns on our service.

These documents are annexed to this report in English, and they are available on our website in all Medi-Cal threshold languages in which we offer Snapchat.

Content moderation policies and practices (Cal. Bus. & Prof. Code, §§22677(a)(3)-(4))

Our Terms prohibit the categories of content referenced in Section 22677(a)(3), as follows:

Category of content referenced in Section 22677(a)	Corresponding category of content prohibited by our Community Guidelines	Relevant definitions and policies, as provided in our Transparency Report Glossary and Community Guidelines explainer series
Hate speech or racism	Hate Speech (which falls under Hateful Content, Terrorism, and Violent Extremism)	Content that demeans, or promotes discrimination towards, an individual or group of individuals on the basis of their race, color, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, veteran status, immigration status, socio-economic status, age, weight, or pregnancy status. For more information, please review our explainer on Hateful Content, Terrorism, and Violent Extremism.
Extremism or radicalization	Terrorism & Violent Extremism (which falls under Hateful Content, Terrorism, and Violent Extremism)	Content that promotes or supports terrorism or other violent, criminal acts committed by individuals and/or groups to further ideological goals, such as those of a political, religious, social, racial, or environmental nature. It includes any content that promotes or supports any foreign terrorist organization or violent extremist hate group, as well as content that advances recruitment for such organizations or violent extremist activities. For more information, please review our explainer on Hateful Content, Terrorism, and Violent Extremism.
Disinformation or misinformation	False Information (which falls under Harmful False or	Includes false or misleading content that causes harm or is malicious, such as denying the existence of

	Deceptive Information)	tragic events, unsubstantiated medical claims, or undermining the integrity of civic processes, or manipulating content for false or misleading purposes. For more information, please review our explainer on Harmful False or Deceptive Information.
Harassment	Harassment & Bullying	Refers to any unwanted behavior that could cause an ordinary person to experience emotional distress, such as verbal abuse, sexual harassment, or unwanted sexual attention. This category also includes the sharing or receipt of non-consensual intimate imagery (NCII). For more information, please review our explainer on Harassment & Bullying.
Foreign political interference	False Information (which falls under Harmful False or Deceptive Information).	For our definition of False Information, please see above. Impersonation occurs when an account is falsely pretending to be associated with another person or brand. For more information, please review our explainer on Harmful False or Deceptive Information.
Controlled substance distribution	Drugs (which falls under Illegal or Regulated Activities)	Refers to distribution and use of illegal drugs (including counterfeit pills), and other illicit activity involving drugs. For more information, please review our explainer on Illegal or Regulated Activities.

Our Moderation, Enforcement and Appeals Explainer and Severe Harm Explainer provide detailed information on, among other topics:

- how we moderate content through both automated tools and human review,
- how we respond to user reports of alleged violations of our Community Guidelines, and
- how we enforce against individual pieces of content and users that violate our Community Guidelines.

Information on Violations of our Terms (July 1 - September 30, 2023) (Cal. Bus. & Prof. Code, §22677(a)(5))

Below we provide detailed information about violations of our Community Guidelines that were either reported to us or automatically detected by our systems in the period July 1 - September 30, 2023, consistent with Section 22677(a). We first provide global figures, followed by U.S. figures. These figures relate not only to the categories of violating content referenced in Section 22677(a)(3), but more broadly to the violations referenced in our Community Guidelines.[1]

Except where otherwise specified, terms used in this section are defined in accordance with our Transparency Glossary.

[1] In this report, we have disaggregated the data into: (i) categories of violating content, (ii) how the content or account was flagged (i.e., by a report or by our automated detection tools), and (iii) how the content or account was enforced (i.e., by human reviewers or by automated tools). We are not able to disaggregate the data per type of content (e.g., posts, comments, messages, user profiles) or per type of media (e.g., text, image, video) at this time, because we were not tracking this data globally or in the United States as of Q3 2023, in a manner that would enable us to extract this data for reporting purposes.

Global figures

Category of violation	Manner Flagged	Total Content or Accounts Flagged[1]	Content Enforced[2] by Human Reviewers	Content Enforced by Automated Tools	Unique Accounts Enforced[3] by Human Reviewers	Unique Accounts Enforced by Automated Tools	Appeals Against Account Locks[4] Enforced by Human Reviewers	Appeals Against Account Locks Enforced by Automated Tools	Accounts Reinstated Following Appeal[5] (Initially Locked by Human Reviewers)	Accounts Reinstated Following Appeal (Initially Locked by Automated Tools)	Violative View Rate (VVR)[6] for Content Enforced by Human Reviewers	VVR for Content Enforced by Automated Tools	Unique Violative Viewer Rate[7] for Content Enforced by Human Reviewers	Unique Violative Viewer Rate for Content Enforced by Automated Tools
Hate Speech	Human Report	189,981	45,028	257	39,567	183	206	5	11	0	0.000193%	0.000001%	0.44%	0.002%
	Automatic Detection	148	148	0	132	0	0	0	0	0	0.000000%	0.000000%	0.00%	0.000%
Terrorism & Violent Extremism	Human Report	41,399	835	24	751	21	17	0	1	0	0.000005%	0.000000%	0.01%	0.000%
	Automatic Detection	11	11	0	11	0	0	0	0	0	0.000000%	0.000000%	0.00%	0.000%
False Information	Human Report	216,219	460	10	445	9	3	0	0	0	0.000005%	0.000000%	0.01%	0.000%
	Automatic Detection	16	16	0	16	0	0	0	0	0	0.000000%	0.000000%	0.00%	0.000%
Impersonation	Human Report	213,879	8,040	36	8,002	33	769	0	51	0	0.000002%	0.000000%	0.01%	0.000%
	Automatic Detection	5	5	0	5	0	0	0	0	0	0.000000%	0.000000%	0.00%	0.000%
Harassment & Bullying	Human Report	4,531,005	505,999	20,239	414,702	11,285	14,546	943	410	13	0.001143%	0.000044%	1.52%	0.051%
	Automatic Detection	2,523	2,481	42	2,268	12	78	3	7	0	0.000002%	0.000000%	0.00%	0.000%
Drugs	Human Report	177,028	115,835	5,010	84,731	4,118	8,331	1,056	231	5	0.000536%	0.000031%	0.75%	0.062%
	Automatic Detection	636,008	286,538	158,894	242,067	128,763	73,446	20,420	1,992	103	0.000101%	0.000010%	0.23%	0.028%
Threats & Violence	Human Report	401,227	44,172	5,210	34,555	3,648	747	4	35	0	0.000678%	0.000035%	1.08%	0.064%
	Automatic Detection	410	323	11	292	6	42	0	0	0	0.000000%	0.000000%	0.00%	0.000%
Self-Harm & Suicide	Human Report	85,339	15,896	56	14,637	33	18	1	5	0	0.000007%	0.000000%	0.01%	0.000%
	Automatic Detection	260	252	0	242	0	2	0	0	0	0.000000%	0.000000%	0.00%	0.000%
Spam	Human Report	1,254,516	311,954	514,111	269,775	312,043	7,287	128	108	1	0.000858%	0.000106%	1.28%	0.218%

	Automatic Detection	50,890	15,636	35,254	14,084	21,633	443	96	7	0	0.000004%	0.000029%	0.01%	0.021%
Weapons	Human Report	48,967	6,129	568	4,831	409	214	45	6	1	0.000035%	0.000001%	0.06%	0.002%
	Automatic Detection	123,755	40,106	66,208	32,953	51,275	612	995	25	8	0.000022%	0.000006%	0.06%	0.016%
Other Regulated Goods	Human Report	228,900	68,618	4,582	52,689	2,351	3,989	508	111	4	0.000526%	0.000018%	0.87%	0.029%
	Automatic Detection	9,967	9,925	42	8,668	21	389	25	27	1	0.000010%	0.000000%	0.03%	0.001%
Sexual Content	Human Report	2,146,825	794,265	398,293	580,110	249,112	60,534	4,233	747	19	0.004442%	0.001858%	3.08%	1.392%
	Automatic Detection	397,538	150,421	194,379	98,190	111,567	11,177	1,392	125	10	0.000061%	0.000011%	0.10%	0.019%
Child Sexual Exploitation	Human Report	389,163	113,454	2,547	96,106	1,949	13,677	68	2,059	11	0.000300%	0.000020%	0.45%	0.017%
	Automatic Detection	168,527	78,427	60,312	54,058	44,284	9,124	9,170	745	2,015	0.000002%	0.000000%	0.00%	0.001%
Totals		11,314,506	2,614,974	1,466,085	1,920,608	910,767	205,651	39,092	6,703	2,191	0.008932%	0.002172%	5.99%	1.694%

U.S. figures

Category of violation	Manner Flagged	Total Content or Accounts Flagged[1]	Content Enforced[2] by Human Reviewers	Content Enforced by Automated Tools	Unique Accounts Enforced[3] by Human Reviewers	Unique Accounts Enforced by Automated Tools	Appeals Against Account Locks[4] Enforced by Human Reviewers	Appeals Against Account Locks Enforced by Automated Tools	Accounts Reinstated Following Appeal[5] (Initially Locked by Human Reviewers)	Accounts Reinstated Following Appeal (Initially Locked by Automated Tools)	Violative View Rate (VVR)[6] for Content Enforced by Human Reviewers	Violative View Rate (VVR) for Content Enforced by Automated Tools	Unique Violative Viewer Rate[7] for Content Enforced by Human Reviewers	Unique Violative Viewer Rate for Content Enforced by Automated Tools
Hate Speech	Human Report	74,256	26,254	184	22,888	127	118	0	7	0	0.0004208%	0.0000048%	1.316%	0.015%
	Automatic Detection	86	86	0	79	0	0	0	0	0	0.0000003%	0.0000000%	0.001%	0.000%
Terrorism & Violent Extremism	Human Report	10,901	197	6	190	4	4	0	0	0	0.0000062%	0.0000001%	0.020%	0.000%
	Automatic Detection	6	6	0	6	0	0	0	0	0	0.0000000%	0.0000000%	0.000%	0.000%
False Information	Human Report	47,421	235	3	223	3	0	0	0	0	0.0000072%	0.0000002%	0.023%	0.001%
	Automatic Detection	10	10	0	10	0	0	0	0	0	0.0000000%	0.0000000%	0.000%	0.000%
Impersonation	Human	54,948	2,461	13	2,442	11	241	0	16	0	0.0000001%	0.0000000%	0.000%	0.000%

	Report													
	Automatic Detection	2	2	0	2	0	0	0	0	0	0.0000000%	0.0000000%	0.000%	0.000%
Harassment & Bullying	Human Report	1,134,660	166,787	4,658	140,939	3,385	3,987	89	173	9	0.0017937%	0.0000227%	4.261%	0.051%
	Automatic Detection	1,189	1,186	3	1,092	2	28	1	4	0	0.0000043%	0.0000000%	0.014%	0.000%
Drugs	Human Report	76,888	54,098	1,922	39,227	1,655	3,439	166	96	0	0.0014146%	0.0000292%	2.821%	0.111%
	Automatic Detection	369,835	170,066	117,427	142,887	93,734	37,681	11,458	909	58	0.0002741%	0.0000334%	0.980%	0.141%
Threats & Violence	Human Report	117,412	16,571	1,432	13,448	1,057	316	0	22	0	0.0006518%	0.0000524%	1.691%	0.134%
	Automatic Detection	222	167	10	153	5	26	0	0	0	0.0000007%	0.0000001%	0.002%	0.000%
Self-Harm & Suicide	Human Report	29,226	8,027	8	7,583	8	6	0	3	0	0.0000126%	0.0000000%	0.040%	0.000%
	Automatic Detection	159	153	0	146	0	0	0	0	0	0.0000000%	0.0000000%	0.000%	0.000%
Spam	Human Report	580,657	137,514	360,649	124,895	223,162	1,997	19	22	0	0.0009065%	0.0000995%	2.147%	0.326%
	Automatic Detection	15,974	6,304	9,670	6,126	6,276	122	1	2	0	0.0000037%	0.0000129%	0.016%	0.030%
Weapons	Human Report	17,212	1,742	80	1,604	72	66	9	3	0	0.0000382%	0.0000009%	0.142%	0.004%
	Automatic Detection	99,084	32,206	56,158	26,788	43,961	449	209	17	4	0.0000886%	0.0000241%	0.345%	0.101%
Other Regulated Goods	Human Report	73,261	13,629	306	11,770	210	340	20	23	1	0.0004930%	0.0000038%	1.482%	0.012%
	Automatic Detection	3,539	3,534	5	3,173	3	63	0	10	0	0.0000098%	0.0000000%	0.041%	0.000%
Sexual Content	Human Report	584,728	221,552	127,380	163,531	84,979	16,924	824	233	8	0.0057545%	0.0025898%	8.864%	4.121%
	Automatic Detection	109,214	39,790	44,859	27,328	29,015	3,926	238	38	5	0.0001110%	0.0000145%	0.327%	0.042%
Child Sexual Exploitation	Human Report	109,155	25,071	245	22,045	181	13,677	68	2,059	11	0.0001473%	0.0000049%	0.290%	0.011%
	Automatic Detection	33,376	12,754	11,707	9,686	8,503	9,124	9,170	745	2,015	0.0000009%	0.0000001%	0.002%	0.000%
Totals		3,543,421	940,402	736,725	725,906	486,592	205,651	39,092	6,703	2,191	0.0121398%	0.0028934%	16.290%	4.772%

[1] Total number of pieces of content or accounts that were flagged for potential violations of our Community Guidelines, including those reported to us and those detected through our automated tools. To disaggregate this data into categories of violative content, we've used the ultimate enforcement reason where an enforcement action was taken. Where the content or account was flagged but no enforcement action was taken, we attribute the metrics to the suspected violation category for which the content or account was flagged.

[2] The number of pieces of content (e.g., Snaps, Stories) that were enforced against on Snapchat. "Enforcement" refers to an action taken against a piece of content or an account (e.g., deletion, warning, locking).

[3] The number of unique accounts that were enforced against on Snapchat. For example, if a single account was enforced against multiple times for various reasons (e.g., a user was warned for posting false information and then later deleted for harassing another user), only one account would be calculated in this metric. As above, "enforcement" refers to an action taken against a piece of content or an account (e.g., deletion, warning, locking).

[4] Users can only submit appeals against an account lock.

[5] We only reinstate accounts that our moderators determine were incorrectly locked.

[6] Violative View Rate is the percentage of Story and Snap views that contained violating content, as a proportion of all Story and Snap views across Snapchat. For example, if our VVR is 0.03%, that means for every 10,000 Snap and Story views on Snapchat, 3 contained content that violated our policies. This metric allows us to understand what percentage of views on Snapchat come from content that violates our Community Guidelines (that was either reported or proactively enforced on).

[7] Unique Violative Viewer Rate is the percentage of unique viewers that saw violating content, as a proportion of unique users active throughout the reporting period, i.e., Q3 2023. For example, if our Unique Violative Viewer Rate is 0.03%, that means that, for every 10,000 active users during the relevant period on Snapchat, 3 viewers saw content that violated our policies. This metric allows us to understand what percentage of users on Snapchat come across content that violates our Community Guidelines (that was either reported or proactively enforced on).

Additional information

Although not required by Section 22677, we also believe it valuable to provide our median turnaround times (TATs) for responding to reports and appeals. We define TAT as the time between when our Trust & Safety teams or Automated Tools first receive a report (usually when a report is submitted or detected via automated means) to the last enforcement action timestamp. If multiple rounds of review occur, the final time is calculated at the last action taken. With that in mind, our Global median TAT for content and account reports is approximately 6 minutes.

For additional information regarding Snap's approach to Safety, Privacy, and Transparency, visit our Privacy & Safety Hub, and our About Transparency Reporting page.