Rom Bar-Nissim (SBN 293356)
Rom@HeahBarNissim.com
HEAH BAR-NISSIM LLP
1801 Century Park East, Suite 2400
Los Angeles, CA 90067
Telephone: (310) 432-2836

Jarrett Lee Ellzey (*pro hac vice application forthcoming*)
Tom Kherkher (*pro hac vice application forthcoming*)
Leigh S. Montgomery (*pro hac vice application forthcoming*)
ELLZEY KHERKHER SANFORD MONTGOMERY LLP
4200 Montrose Street, Suite 200
Houston, TX 77006
JEllzey@EKSM.com
TKherkher@EKSM.com
LMontgomery@EKSM.com

Attorneys for Plaintiffs and the Proposed Class

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

TED ENTERTAINMENT, INC., MATT FISHER, and GOLFHOLICS, INC., each individually and on behalf of all others similarly situated,

Plaintiffs,

v.

OPENAI, INC., OPENAI GP LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OpenAI Group PBC, and OPENAI HOLDINGS, LLC

Defendants

Case No.: 3:26-cv-2935

CLASS ACTION COMPLAINT

DEMAND FOR JURY TRIAL

Plaintiffs Ted Entertainment, Inc. ("TEI"), Matt Fisher, and Golfholics, Inc. (collectively, where appropriate "Plaintiffs"), each individually and on behalf of all others similarly situated (the "Class," as defined below), by and through their undersigned counsel, file this Complaint against OpenAI, Inc.; OpenAI GP LLC; OpenAI OpCo LLC; OAI Corporation, LLC; OpenAI Holdings, LLC; and OpenAI Group PBC (collectively, "OpenAI" or "Defendant"), and alleges as follows:

NATURE OF THE ACTION

1. This lawsuit arises from Defendant unlawfully circumventing technological protection measures to access and scrape millions of copyrighted videos from the online video viewing platform, YouTube, in order to feed, train, improve, and commercialize Defendant's large-scale generative artificial intelligence ("AI") model named "Sora."

2. Sora is a text-to-video AI model conceived and marketed as OpenAI's dedicated video generation platform. Sora was significant enough that Defendant launched it as its own standalone application, the Sora app, separate from and in addition to its integration within ChatGPT.

3. YouTube allows the public to view audiovisual works only through controlled streaming, keeping its underlying video files locked away behind careful access restrictions. Defendant picked this lock by deploying scraping tools designed to evade and circumvent YouTube's access restrictions. Defendant's circumvention of YouTube's technological protection measures ("TPMs") was systematic. On information and belief, Defendant used automated video-downloading programs combined with virtual machines that rotated IP addresses to avoid detection and blocking, enabling the mass unauthorized access and extraction of videos at the scale necessary to train Sora. Defendant then used Plaintiffs' and Class Members' intellectual property for their own commercial gain in developing and commercializing Sora. In doing so, Defendant has violated the law and YouTube's Terms of Service, which were intended to protect Plaintiffs and others similarly situated.

4. By accessing, scraping, and downloading those files, Defendant deliberately circumvented YouTube's access controls to obtain the training data necessary to build and commercialize Sora.

5. Plaintiffs and the Class Members whom Plaintiffs seek to represent are content creators who upload their audiovisual content to YouTube. In doing so, the content creators are authorizing and instructing YouTube to provide protection to the video content through YouTube's anti-circumvention software and Terms of Service. In fact, YouTube's anti-circumvention software and protective Terms of Service are a driving factor behind many content creators' decision to upload their video content to YouTube.

6. Rather than seek permission or pay a fair price for the audiovisual content hosted on YouTube, Defendant harvested content creators' protected and copyrighted videos for commercial use and at scale without consent or compensation to the content creators.

7. Defendant's actions were not only unlawful, but an unconscionable attack on the community of content creators whose content is used to fuel the multi-trillion-dollar generative AI industry without any compensation.

8. Content creators such as Plaintiffs and the Class Members will never be able to claw back the intellectual property unlawfully copied and used by Defendant to train its generative AI models. Once AI ingests content, that content is stored in its neural network and not capable of deletion or retraction. Defendant's actions constitute abuse and exploitation of content creators' work for Defendant's profit.

9. Most YouTube videos are not registered with the U.S. Copyright Office. That lack of registration, however, does not render them valueless or leave them unprotected. Content creators invest time, skill, and resources into producing their works, and they rely on YouTube's TPMs to safeguard their files from unauthorized access. Because copyright registration is not a prerequisite for protection against unlawful circumvention of access controls, this claim is especially critical where Defendant's misconduct lies in breaking through access barriers that prevent anyone from obtaining the underlying files in the first place.

10. It is also critical to protect a flourishing internet ecosystem. In a world where Defendant and others can circumvent technological protections to exploit copyrighted works without authorization with impunity, creators will be less likely to make their creations available on YouTube and other similar platforms, for fear of losing all control of them. The world will be poorer

for it.

11. The Class Members, including Plaintiffs, are content creators who uploaded their audiovisual content to YouTube's video platform. On information and belief, the audiovisual content of Class Members was among the video content utilized by Defendant to train Sora and related generative AI models.

12. An essential component of Defendant's business model—powering AI features and services with large-scale training data—includes the mass acquisition and ingestion of creators' videos scraped from YouTube.

13. Defendant has profited substantially from its infringement of Plaintiffs' and Class Members' video content through Defendant's training of its generative AI products. Defendant's massive financial success would not have been possible without the video content created by Plaintiffs and Class Members, which was intended for streaming on YouTube.

14. Plaintiffs bring this class action on behalf of themselves and on behalf of a nationwide class of YouTube creators whose works were scraped, ingested, and trained on without authorization, seeking statutory damages, injunctive relief, restitution, and all other remedies allowed by law pursuant to the Digital Millennium Copyright Act, 17 U.S.C.A. § 1201(a) seeking an injunction and damages commensurate with the scope of Defendant's massive and ongoing infringement. More particularly, Defendant's conduct violates the provisions of the DMCA regarding anti-circumvention (§1201) by bypassing technological protection measures that control access to YouTube videos.

THE PARTIES

15. Plaintiff Ted Entertainment, Inc. ("TEI"), is an independent California based media company and content creator with over 5,800 original videos on YouTube, with a combined total of over 4 billion views. TEI has amassed a substantial following on YouTube with over 2.6 million subscribers to its channels. TEI owns and operates the channels "h3h3 Productions" and "H3 Podcast Highlights." H3H3 Productions appears with 146 videos in HD-VILA-100M, 83 videos in HD-VG-130M, 155 videos in Panda-70M, and 24 videos in HowTo100M. H3 Podcast Highlights appears with 285 videos in HD-VILA-100M, 209 videos in HD-VG-130M, 283 videos in Panda-70M, and

1 video in HowTo100M. Plaintiff invested significant resources – time and money – into producing and publishing this video content and bringing awareness to its content.

16. Plaintiff TEI is the creator of the videos identified in Exhibit A, all of which were uploaded exclusively to YouTube by Plaintiff. Plaintiff derives value from the works identified in Exhibit A through viewership, advertising, sponsorships, licensing, and related monetization.

17. Plaintiff TEI and its owners – Ethan and Hila Klein – are longtime champions of the rights of YouTube creators. The Kleins helped define what constitutes fair use reaction videos and the good faith belief standard for DMCA counternotifications. TEI has championed online free speech and is currently helping define what reaction content does not constitute fair use to ensure YouTube content creators can enjoy the fruits of their labor.

18. Plaintiff Matt Fisher is an individual and resident of the State of California. He is a golf content creator who posts his original videos on YouTube, many of which are instructional. His channel is "MrShortGame Golf" on the YouTube platform. He has over 500,000 subscribers and hundreds of millions of views. Plaintiff has invested substantial time and money into bringing awareness around his content. MrShort Game appears with 2 videos in HD-VILA-100M, 1 video in HD-VG-130M, 8 videos in Panda-70M, and 4 videos in HowTo100M.

19. Plaintiff Matt Fisher is the creator of the videos identified in Exhibit B, all of which were uploaded exclusively to YouTube by Plaintiff and all of which were included in datasets used by Defendant to train its generative AI models. Plaintiff derives value from the works identified in Exhibit B through viewership, advertising, sponsorships, licensing, and related monetization.

20. Plaintiff Golfholics is a corporate entity organized pursuant to the laws of the State of California. It is a golf content channel that posted its original videos on YouTube. The channel is "Golfholics" on the YouTube platform. It has over 130,000 subscribers and millions of views. Plaintiff invested substantial time and money into bringing awareness around its content. Golfholics appears with 62 videos in HD-VILA-100M, 62 videos in Panda-70M, and 37 video in HD-VG-130M.

21. Plaintiff Golfholics is the creator of the videos identified in Exhibit C, all of which were uploaded exclusively to YouTube by Plaintiff and all of which were included in datasets used

by Defendant to train its generative AI models. Plaintiff derives value from the works identified in Exhibit C through viewership, advertising, sponsorships, licensing, and related monetization.

22. Defendants are closely related Delaware entities that together operate as OpenAI. Defendant OpenAI Inc. is a Delaware nonprofit corporation with a principal place of business located in San Francisco, California. OpenAI Inc. indirectly owns and controls all other OpenAI entities, including all other Defendants, and has been directly involved in carrying out the large-scale 17 U.S.C. §1201(a) anti-circumvention violations alleged in this Complaint.

23. Defendant OpenAI GP, LLC is a Delaware limited liability company with a principal place of business located in San Francisco, California. OpenAI GP LLC is wholly owned and controlled by OpenAI Inc. OpenAI GP, LLC has been directly involved in carrying out the large-scale 17 U.S.C. §1201(a) anti-circumvention violations alleged in this Complaint through its direction and control of OpenAI LP and OpenAI Global LLC.

24. Defendant OpenAI Global, LLC is a Delaware limited liability company with a principal place of business located in San Francisco, California. OpenAI Global LLC owns, sells, licenses, and monetizes a number of OpenAI's offerings, including ChatGPT, ChatGPT Enterprise, and OpenAI's API tools, all of which were built on OpenAI's large-scale 17 U.S.C. §1201(a) anti-circumvention violations alleged in this Complaint. Upon information and belief, OpenAI Global, LLC is owned and controlled by OpenAI Inc.

25. Defendant OpenAI OpCo LLC is a Delaware limited liability company with a principal place of business located in San Francisco, California. OpenAI OpCo LLC is a wholly owned subsidiary of OpenAI Inc. and has facilitated and directed OpenAI's largescale 17 U.S.C. §1201(a) anti-circumvention violations alleged in this Complaint through its management and direction of OpenAI Global, LLC.

26. Defendant OAI Corporation, LLC is a Delaware limited liability company with a principal place of business located in San Francisco, California. OAI Corporation, LLC's sole member is OpenAI Holdings, LLC/OpenAI Group PBC. OAI Corporation, LLC was and is involved in the large-scale 17 U.S.C. §1201(a) anti-circumvention violations alleged in this Complaint through its ownership, control, and direction of OpenAI Global LLC and OpenAI LLC.

27. Defendant OpenAI Holdings, LLC is a Delaware limited liability company with a principal place of business located in San Francisco, California whose sole members are OpenAI, Inc. and Aestas, LLC, whose sole member, in turn, is Aestas Management Company, LLC. Aestas Management Company, LLC is a Delaware shell company that OpenAI used to execute a $495 million capital raise in 2023. According to OpenAI, Defendant OpenAI Holdings, LLC merged into OpenAI Group PBC effective December 31, 2025.

28. OpenAI Group PBC is a Delaware benefit corporation with a principal place of business located in San Francisco, California. OpenAI Group PBC was incorporated in October 2025. According to OpenAI, OpenAI Holdings, LLC merged into OpenAI Group PBC effective December 31, 2025. Through that merger, OpenAI Group PBC is the successor to OpenAI Holdings, LLC's conduct as alleged in this Complaint, and is liable for the harm caused by the conduct of its predecessor entities.

JURISDICTION AND VENUE

29. This is a civil action seeking damages and injunctive relief for infringement under the Digital Millennium Copyright Act, 17 U.S.C. §§ 1201–1205. As such, this Court has subject matter jurisdiction over this action pursuant to 28 U.S.C. §§ 1331 and 1338(a), based on federal question jurisdiction.

30. The Court has supplemental jurisdiction over related state-law claims, if any, under 28 U.S.C. § 1367.

31. Personal jurisdiction is proper because Defendant's principal place of business is in San Francisco, California, transacts business nationwide, purposefully avails itself of this forum, and a substantial part of the events or omissions giving rise to these claims occurred or were directed here.

32. Venue lies in this judicial district pursuant to 28 U.S.C. § 1391(b)(1) and/or 1400(a) because OpenAI resides in, is found in, or may be sued in this District, and a substantial part of the events giving rise to the claims occurred here.

/ / /

/ / /

FACTUAL BACKGROUND

A. YouTube's Terms of Service Prohibit Scraping, Downloading, and Unauthorized Access to Video Files, and Employs Technological Measures Consistent With This Prohibition

33. As noted above, Plaintiffs and the Class Members are content creators of audiovisual content.

34. Plaintiffs and the Class Members create original video content that has been uploaded onto YouTube's video sharing platform.

35. YouTube's "users"—the individuals who view the digital content available on YouTube—can watch and listen to videos for free on YouTube's advertising-supported service, but YouTube does not give users access to or allow downloading of the digital files underlying the content viewed by the user through YouTube's authorized playback mechanism.

36. YouTube's Terms of Service expressly prohibit scraping, unauthorized downloading, bulk extraction, or other forms of data mining of audiovisual content except through expressly permitted features or licensed APIs. These contractual restrictions operate together with YouTube's TPMs to prevent unlicensed access to creators' videos.

37. According to YouTube's Terms of Service, content creators such as Plaintiffs who upload content onto YouTube grant license to YouTube for certain uses as well as to other users of YouTube to access content through YouTube's services; however, the license makes clear that it "does not grant any rights or permissions for a user to make use of [the] Content independent of the Service."[1]

38. This language confirms that end users are not given access to the file itself, only the ability to view (*i.e.*, stream) through YouTube's controlled environment. YouTube's Terms of Service reflect YouTube's intent to restrict access to the digital files underlying the videos that YouTube's users are allowed to stream through YouTube's platform.

39. Streaming through YouTube and downloading permanent copies provide the user with different value propositions—watching and listening for free but seeing ads, versus possessing

[1] "Terms of Service," YouTube, https://www.youtube.com/t/terms (last viewed March 11, 2026).

a permanent digital copy.

40. Accordingly, scraping or bulk downloading is not merely copying material already provided; it is an act of unauthorized access to data files that YouTube affirmatively withholds from public download.

41. In fact, during a Bloomberg interview about OpenAI scraping YouTube's CEO, Neal Mohan, confirmed that unauthorized scraping constitutes a violation of creators' rights and YouTubes Terms of Service stating, "**From a creator's perspective, when a creator uploads their hard work to our platform, they have certain expectations. One of those expectations is that the terms of service is going to be abided by**," Mohan said, "**It does not allow for things like transcripts or video bits to be downloaded, and that is a clear violation of our terms of service. Those are the rules of the road in terms of content on our platform.**"[2]

42. The scraping and acquisition processes used by Defendant to circumvent YouTube's TPMs were inconsistent with, and in violation of, YouTube's Terms of Service, which forbid scraping and mass downloading of videos.[3]

43. To enforce its prohibitions on downloading content, YouTube does not provide a downloading option that is readily available to users.

44. Although YouTube offers downloading options to subscribers who pay for YouTube's "Premium" plan, YouTube still employs TPMs to restrict access. Specifically, YouTube prohibits all downloading audiovisual content except to the YouTube app. Further, the "download" option only makes audiovisual content available for offline streaming—it does not provide the Premium subscriber with access to the audiovisual files. Accordingly, the audiovisual files cannot be transferred to any other device, but remain only for streaming on the app. Finally, the files are available only for offline streaming for a limited amount of time, at which point they will become available only for online streaming once again.

[2] Davey Alba & Emily Chang, *YouTube Says OpenAI Training Sora With Its Videos Would Break Rules*, Bloomberg (Apr. 4, 2024), https://www.bloomberg.com/news/articles/2024-04-04/youtube-says-openai-training-sora-with-its-videos-would-break-the-rules (video available at https://www.youtube.com/watch?v=FBZ__BeChRg).

[3] *Supra*, note 1 (prohibiting "access, reproduce, download, distribute, transmit, broadcast, display, sell, license, alter, modify or otherwise use any part of the Service or any Content except: (a) as expressly authorized by the Service; or (b) with prior written permission from YouTube and, if applicable, the respective rights holders;").

45. For users who do not have a "Premium" plan, the "download" option on YouTube's player page is non-functional.

46. To further enforce its prohibitions on downloading content, YouTube uses technological processes and tools to detect and block access to files for unauthorized downloading. For example, YouTube monitors downloading activity and may block IP addresses that make too many download attempts in a specified period.

47. YouTube has employed a variety of TPMs to restrict access to Class Members' audiovisual files. For example, YouTube does not provide users with a readily available downloading option and it has implemented a number of tools, such as the Rolling Cipher and IP Blocking, that operate to restrict users' ability to access audiovisual material.

48. Content creators, including Plaintiffs and Class Members, rely on the TPMs and YouTube's Terms of Service in deciding to upload their audiovisual content to YouTube. Content creators, including Plaintiffs and Class Members, expect that their works will not be copied at scale without consent.

B. YouTube's Technological Protection Measures Function as Access Controls Over Video Files.

49. YouTube deploys multiple technological protection measures ("TPMs") designed to control, restrict, and monitor access to the underlying video files and to deter direct downloading or bulk extraction outside YouTube's controlled playback environment. These TPMs include, among others: (1) an obfuscated signature system commonly referred to as a "rolling cipher," (2) IP-based blocking and rate limiting that restrict high-volume automated access, (3) short-lived, session-bound streaming URLs, (4) CAPTCHA human-verification challenges triggered by automated activity, and (5) proof-of-origin tokens that verify requests originating from authorized client environments. Each TPM independently functions as a gatekeeping mechanism that must be satisfied before the audiovisual file can be retrieved, and circumvention of any one of them enables access to the file in a manner not authorized by YouTube or the content owners.

50. **TPM (1) - Rolling cipher (access control):** YouTube controls access to the underlying media file by withholding a usable file location unless the requesting client can transform

an obfuscated signature parameter using proprietary logic embedded in the official YouTube player. The playback data delivered to users contains a scrambled signature that must be processed to produce a valid media request URL. Without performing this authorized transformation, the server will not deliver the audiovisual file. Thus, the rolling cipher controls access by requiring application of a specific computational process before the file can be obtained at all.

51. YouTube's rolling cipher encryption measure acts as a digital lock, controlling access to content by protecting against unauthorized downloading of underlying media files. YouTube maintains two different URLs for any given video: the page URL, visible to the user, is for the webpage where the video playback occurs, and the file URL, not visible to the user, is for the video file itself that is played within the page. The file URL is encrypted using a complex and periodically changing algorithm – the rolling cipher – that is designed to impede external access to the underlying YouTube files. YouTube's player software uses a decryption routine to authenticate requests and deliver content only through approved interfaces. This TPM inhibits access to the underlying audiovisual files for the purposes of any downloading, copying or distribution of the audiovisual content. In other words, the rolling cipher controls access to copies of audiovisual content uploaded to the platform and impedes an ordinary user from creating a permanent, unrestricted download of audiovisual content made available on YouTube only for streaming and restricted downloading.

52. The rolling cipher is a TPM measure within the meaning of 17 U.S.C. §1201(a) because it "effectively controls access" to copyrighted works by preventing users access to the files of video streams without first executing YouTube's proprietary decryption code.

53. **TPM (2) - IP blocking and rate limiting (access control):** YouTube monitors network behavior and restricts access when excessive or abusive request patterns are detected. Requests from an IP address that exceeds defined thresholds may be throttled, denied, or blocked entirely, preventing further retrieval of audiovisual data from that source. This mechanism controls access by limiting how frequently and at what scale a particular requester may obtain the file and by cutting off access altogether when abuse is detected.

54. YouTube's infrastructure monitors the number and frequency of video requests

originating from individual IP addresses. This system serves an important function. Even if a user were able to generate a valid media request for an individual video, YouTube's infrastructure is designed to prevent automated systems from retrieving large numbers of videos at scale.

55. Once YouTube detects abnormal volume of access to files, the system can block the offending IP address from accessing YouTube's servers. This prevents automated scraping systems from accessing large volumes of copyrighted audiovisual content.

56. This IP-based monitoring and blocking system therefore also functions as a TPM measure within the meaning of 17 U.S.C. §1201(a) because it "effectively controls access" to copyrighted works by restricting automated access to YouTube's underlying video files.

57. **TPM (3) - Session-bound, short-lived URLs (access control):** Even after a valid media URL is generated, YouTube restricts access by issuing URLs that are temporary, cryptographically signed, and tied to a particular playback session and client context. These URLs include authorization parameters such as expiration timestamps and client identifiers. Once the authorization window expires or the session ends, the server will refuse to deliver the audiovisual file in response to that URL. Accordingly, possession of a previously valid link does not provide continuing access to the file, and new authorization must be obtained to retrieve it. This mechanism therefore controls access by limiting when and from where the file may be requested.

58. Because session-bound URLs expire automatically, any automated system attempting to access videos outside ordinary playback cannot rely on a static link and must instead repeatedly obtain fresh authorization parameters and regenerate valid media URLs. By programmatically renewing expired credentials and initiating new authorized sessions at machine speed, automated scraping tools can maintain continuous access to audiovisual files that would otherwise become unavailable once the original authorization lapses. This conduct circumvents the temporal and session-based restrictions imposed by YouTube and allows persistent retrieval of files in a manner not available to ordinary users. Circumventing this measure constitutes circumvention of a technological measure that effectively controls access to a copyrighted work in violation of 17 U.S.C. § 1201(a).

59. **TPM (4) - CAPTCHA challenges (access control):** When traffic patterns indicate

automated or suspicious activity, YouTube may require completion of a CAPTCHA challenge before allowing further requests to proceed. Until the challenge is successfully completed, the requesting client cannot obtain the data necessary to retrieve the audiovisual file. CAPTCHAs therefore function as an access control by conditioning continued access to content on proof that the requester is a human user rather than an automated system.

60. Large-scale scraping operations necessarily generate request patterns that would ordinarily trigger such human-verification challenges. To continue accessing without interruption, automated systems must be configured either to evade detection, distribute requests across multiple machines, or otherwise prevent CAPTCHA enforcement from halting access. By avoiding or neutralizing these verification gates, such systems obtain audiovisual data without demonstrating the human interaction that YouTube requires for continued access. Defeating or bypassing this human-verification mechanism constitutes circumvention of a technological measure that effectively controls access to copyrighted works in violation of 17 U.S.C. § 1201(a).

61. **TPM (5) - Proof-of-origin tokens (access control):** YouTube further restricts access by requiring that requests for video segments include cryptographic tokens demonstrating that the request originates from an authorized client environment operating within the intended playback context. These tokens are generated during playback and validated by YouTube's servers before data is delivered. Requests lacking valid tokens, or presenting forged or replayed tokens, are denied or degraded. This mechanism controls access by preventing unauthorized software from directly requesting the audiovisual file unless it can present credentials proving it is an approved client.

62. Automated tools operate outside YouTube's authorized player environment and therefore must extract, replicate, or reuse the necessary request parameters in order to obtain access to video data. By presenting such credentials outside their intended context, these tools can retrieve audiovisual files without operating as approved clients. This conduct bypasses the origin-verification function of the token system and enables direct access to media files that would otherwise be delivered only to authorized playback software. Such bypassing constitutes circumvention of a technological measure that effectively controls access to copyrighted works in

violation of 17 U.S.C. § 1201(a).

C. Defendant Improperly Circumvented YouTube's Technological Protection Measures to Obtain Millions of YouTube Videos to Train Its Foundational AI Model.

63. As noted above, Defendant require significant amounts of data to feed, train, improve, and commercialize Sora.

64. The full extent of Defendant's unauthorized access and extraction of video content to train Sora is not yet known and will be the subject of discovery. Defendant has compounded this uncertainty by deliberately concealing the origins of Sora's training data behind the vague label of "publicly available" data. When OpenAI CTO Mira Murati was directly asked what data was used to train Sora, her answer was: "We used publicly available data…"[4] When pressed further on whether OpenAI used YouTube videos, Murati struggled to answer and repeatedly refused to clarify. *Id*. This evasion was deliberate. When OpenAI COO Brad Lightcap was later given another opportunity to clarify if YouTube videos were used to train Sora, OpenAI again refused to answer, opting for silence shrouded in a heap of words.[5]

65. The label "publicly available" does not mean lawfully obtained. The fact that a video can be accessed and viewed by the public through a web browser does not mean it can be accessed, downloaded, extracted, and fed into a commercial AI training pipeline. YouTube's content is publicly viewable precisely because YouTube streams it under strict access controls that prohibit downloading. YouTube CEO Neal Mohan made this distinction explicit, stating that YouTube's terms of service "does not allow for things like transcripts or video bits to be downloaded, and that is a clear violation of our terms of service."[6] Defendant cannot hide behind the word "public" to

[4] Joanna Stern, *OpenAI CTO Mira Murati on Sora, Generative Video*, Wall St. J. (Mar. 2024), https://www.wsj.com/tech/personal-tech/openai-cto-sora-generative-video-interview-b66320bb (when asked directly whether Sora was trained on YouTube videos, Murati stated only that Defendants used "publicly available" and declined to confirm or deny the use of any specific platform's content).

[5] Jaron Schneider, *OpenAI Again Refuses to Say if It Used Your Content to Train Sora*, PetaPixel (May 10, 2024), https://petapixel.com/2024/05/10/openai-again-refuses-to-say-if-it-used-your-content-to-train-sora/

[6] Davey Alba & Emily Chang, *YouTube Says OpenAI Training Sora With Its Videos Would Break Rules*, Bloomberg (Apr. 4, 2024), https://www.bloomberg.com/news/articles/2024-04-04/youtube-says-openai-training-sora-with-its-videos-would-break-the-rules (video available at https://www.youtube.com/watch?v=FBZ__BeChRg).

launder what was in reality an unauthorized mass unauthorized access and extraction of protected content.

66. Sora is not a research tool. It is Defendant's commercialized text-to-video generative AI model, monetized through the Sora app and integrated across Defendant's broader product ecosystem. Because Sora was commercialized, Defendant had an overwhelming incentive to acquire training data on an unprecedented scale. Rather than negotiate for lawful licenses, Defendant broke through YouTube's access protections to obtain the underlying video files necessary to fuel Sora and, by extension, Defendant's broader product lines.

67. Although Defendant has refused to identify their training data sources with specificity, their own Sora System Card provides a critical admission. As Defendant stated: "Sora was trained on diverse datasets, including a mix of publicly available data, proprietary data accessed through partnerships, and custom datasets developed in-house. These consist of: Select publicly available data, mostly collected from industry-standard machine learning datasets and web crawls."[7] Defendant thus admits that Sora was trained on what the AI industry recognizes as "industry-standard machine learning datasets."

68. That admission is significant because the field of video-based AI model training has a well-established set of industry-standard datasets. On information and belief, among the industry-standard machine learning datasets on which Defendant trained Sora were HD-VG-130M, HD-VILA-100M, Panda-70M, and HowTo100M. Each of these datasets consists entirely of references and location identifiers pointing to millions of YouTube videos. None contains the underlying audiovisual files themselves. Each therefore required Defendant to access and download the referenced videos directly from YouTube. Meaning that using these datasets for commercial AI training necessarily required the mass circumvention of YouTube's TPMs.

69. HD-VG-130M, HD-VILA-100M, Panda-70M and HowTo100M are not obscure or experimental datasets. They are among the most widely cited and broadly adopted video datasets in the field of AI model development. HowTo100M, published by Miech et al., introduced a large-

[7] OpenAI, *Sora System Card*, OpenAI (Dec. 2024), https://openai.com/index/sora-system-card/ (disclosing that Sora was trained on "a mix of publicly available data" consisting of "select publicly available data, mostly collected from industry-standard machine learning datasets and web crawls" without identifying the specific sources of publicly available data).

scale dataset of 136 million video clips sourced from 1.22 million narrated instructional YouTube videos and demonstrated that text-video embeddings trained on this data produced state-of-the-art results across multiple AI benchmarks.[8] HD-VILA-100M, published by researchers at Microsoft Research Asia, described itself as the largest and most diversified high-resolution video-language dataset available, comprising 100 million clip-sentence pairs drawn from 3.3 million YouTube videos across 15 popular content categories.[9] HD-VG-130M followed as a large-scale dataset compiled specifically for text-to-video generation research and released publicly on GitHub.[10] Panda-70M is a collection of 3.8 million videos from YouTube split into approximately 70.7 million clips and paired with text captions. It was compiled by Snap and released in 2024.[11] All four datasets are routinely cited across the AI research literature as foundational resources for training video generation models, and all three are built substantially from YouTube content.

70. The centrality of these four datasets to the field of video AI is not a matter of dispute. Independent researchers, academic institutions, and AI companies around the world have cited and built upon HowTo100M, HD-VILA-100M, Panda-70M and HD-VG-130M as standard reference points for years. They are, in the plain meaning of the term Defendant used in their Sora System Card, "industry-standard machine learning datasets." When Defendant admitted that Sora was trained on industry-standard machine learning datasets collected from the web, they admitted that Sora was trained on datasets of precisely this kind. The logical inference is direct and supported by the evidence: Defendant used these datasets, built from millions of YouTube videos belonging to Plaintiffs and Class Members, to train Sora commercially, in violation of the express restrictions governing each dataset and in circumvention of YouTube's TPMs.

71. At the time OpenAI developed and trained Sora, the use of Panda-70M and HD-

[8] Antoine Miech et al., *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*, arXiv:1906.03327 (June 7, 2019), https://arxiv.org/abs/1906.03327 (introducing a large-scale dataset of 136 million video clips sourced from 1.22 million narrated instructional YouTube videos spanning over 23,000 visual tasks).

[9] Hongwei Xue et al., *Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions*, arXiv:2111.10337 (Nov. 19, 2021), https://arxiv.org/abs/2111.10337.

[10] Wenjing Wang et al., *VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation*, arXiv:2305.10874 (May 2023), https://arxiv.org/abs/2305.10874 (introducing the HD-VG-130M dataset, consisting of 130 million text-video pairs collected from YouTube in high-definition, widescreen, and watermark-free formats for text-to-video generation research).

[11] Tsai-Shien Chen et al., *Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers*, arXiv:2402.19479 (Feb. 29, 2024), https://arxiv.org/abs/2402.19479.

VG-130M represented the industry standard for large-scale text-to-video model training. This is demonstrated by the independent conduct of multiple research teams who, lacking access to OpenAI's proprietary pipeline, built Sora-equivalent models using precisely these datasets. The PKU-YuanGroup's Open-Sora-Plan, initiated by Peking University researchers explicitly to reproduce Sora, trained its models primarily on Panda-70M across every version of the project, treating it as the foundational video corpus for achieving Sora-class output quality.[12]

72. In another example of an independent research group attempting to recreate Sora, the "Open-Sora" project likewise trained on Panda-70M and HD-VG-130M, identifying these datasets as core components of the open-source training corpus necessary to reproduce Sora's capabilities.[13] The convergent, independent reliance on these specific datasets by multiple sophisticated research teams confirms that they were not merely one option among many. These are datasets a developer would use to build a model like Sora. A company with OpenAI's resources building a model of Sora's capability would have used, at minimum, the same datasets the open-source community identified as necessary, and in all likelihood far more of them.

73. YouTube allows users to stream content but prohibits, through contractual terms and TPMs, accessing the underlying video file of the content it streams. Defendant obtained unauthorized access to audiovisual content by unlawfully accessing and stream ripping, converting streaming content into permanent, locally stored files, and by circumventing the TPMs specifically designed to prevent such access.

74. Each of the four datasets functions as a map or index file identifying specific YouTube videos and clips by URL, video identifier, and timestamp. A single YouTube video may be divided into numerous clips, each treated as a separate training sample. Extracting any clip requires independently accessing the source video on YouTube and isolating the designated segment, a process that constitutes a separate act of circumvention for each clip retrieved.

75. Defendant obtained unauthorized access to audiovisual content to train its AI models by unlawfully "stream ripping" them—*i.e.*, converting the audiovisual streaming content into

[12] Bin Lin et al., *Open-Sora Plan: Open-Source Large Video Generation Model*, arXiv:2412.00131 (Nov. 28, 2024), https://arxiv.org/abs/2412.00131.
[13] Zangwei Zheng et al., *Open-Sora: Democratizing Efficient Video Production for All*, arXiv:2412.20404 (Dec. 29, 2024), https://arxiv.org/abs/2412.20404.

permanent, locally-stored files—from YouTube and circumventing TPMs specifically designed to prevent access for such use.

76. The HD-VG-130M dataset contains references to more than 130 million video clips derived from 1,549,408 YouTube videos. The dataset was created by partitioning approximately 1,549,408 YouTube videos into short high-definition segments intended for machine learning research. Each listed clip corresponds to a specific portion of a source video that must be independently retrieved in order to obtain usable audiovisual data.

77. The compilers of HD-VG-130M selected videos based on automated criteria, including aesthetic scoring, and excluded videos displaying creator names or visible watermarks. The dataset was released for research purposes and does not itself distribute the underlying videos or grant any license to use those works commercially.

78. Defendant used HD-VG-130M to retrieve every clip and incorporate those files into its commercial training pipeline. Because each clip corresponds to a separate act of circumvention, the ingestion of HD-VG-130M required roughly 130 million unauthorized downloads, each constituting a separate circumvention event.

79. Microsoft created a document identifying the content of the HD-VG-130M dataset. Microsoft then released this list into the broader AI ecosystem. Not only did Microsoft assemble the HD-VG-130M dataset internally, but it also posted the dataset publicly on GitHub—a code-sharing platform owned by Microsoft—at https://github.com/daooshee/HD-VG-130M.

80. The GitHub repository includes a license agreement that restricts the dataset to "academic use only," acknowledging that any use for commercial purposes would be unauthorized.

81. In addition, the GitHub license agreement prohibits distribution, reproduction, modification, or exploitation of the dataset content without the permission of the copyright owner, thus recognizing that copyright ownership resides with the original creators of the source material, including YouTube content creators such as Plaintiffs and the Class Members.

82. By including these limitations on its GitHub repository, Microsoft implicitly acknowledged that the HD-VG-130M dataset is comprised of protected, copyrighted works obtained without license, permission, or authorization.

83. On information and belief, Defendant obtained the dataset that was earmarked for academic use only, including the HD-VG-130M dataset. Defendant was aware that the datasets were for research purposes only but used the datasets for commercial development (*i.e.*, "training") of its generative AI model.

84. Further, Defendant used the HD-VG-130M dataset to scrape audiovisual content from YouTube without the consent of YouTube or the content creators.

85. The HD-VILA-100M dataset is comprised of references to roughly 100 million clips that were derived from 3,098,462 YouTube videos. Each clip represents a distinct segment of the original work. The dataset was compiled by Microsoft Research Asia and published in 2021 for research involving video-based AI models. It lists pointers to YouTube videos and includes subtitle information extracted from YouTube's closed captioning system.

86. The usage license on the dataset states it was created "solely for Computational Use for non-commercial research. This restriction means that you may engage in non-commercial research activities (including non-commercial research undertaken by or funded via a commercial entity), but you may not use the Data or any Results in any commercial offering, including as part of a product or service (or to improve any product or service) you use or provide to others."

87. HD-VILA-100M does not contain the video files themselves. Any user of the dataset must download each referenced clip directly from YouTube. Defendant did so at scale, which triggered roughly 100 million separate acts of unauthorized access and copying.

88. By including these limitations on its usage license, Microsoft implicitly acknowledged that the HD-VILA-100M dataset is comprised of protected, copyrighted works obtained without license, permission, or authorization.

89. Defendant obtained the dataset that was earmarked for non-commercial use only, including the HD-VILA-100M dataset. Defendant was aware that the datasets were for research purposes only but used the datasets for commercial development (*i.e.*, "training") of its generative AI model. Further, Defendant used the HD-VILA-100M dataset to scrape audiovisual content from YouTube without the consent of YouTube or the creators of the audiovisual content.

90. The HowTo100M dataset consists of references to more than 100 million clips that

were derived from 1,238,911 YouTube videos. The clips depict 23,611 different tasks related to cooking, handcrafting, personal care, gardening, and other activities. The task labels and descriptions were generated using the taxonomy developed by WikiHow. The dataset is hosted by the European research institutions that created it. Like the others, it contains only identifiers and metadata, not any audiovisual material. To use HowTo100M in training, a company must download each clip directly from YouTube. Defendant carried out these downloads as part of its Sora training pipeline, resulting in more than 100 million additional unauthorized copying events.

91. Although the HowTo100M dataset is posted online with what its creators label as a "commercial license," that license has no effect on the legal rights associated with the underlying YouTube videos. The dataset creators do not own the videos and therefore cannot grant any license that authorizes the reproduction, downloading, or commercial use of those works. The dataset license applies only to the dataset itself, meaning the text files, identifiers, and metadata created by the researchers. It does not and cannot confer any right to copy, download, or redistribute the millions of YouTube videos identified in the dataset. Each of those videos remains protected by YouTube's Terms of Service and by technological access restrictions. As a result, every act of downloading a HowTo100M clip from YouTube is a separate unauthorized accessing and copying event and a separate circumvention violation, regardless of any license attached to the dataset file.

92. Panda-70M is a derivative dataset created by Snap that used the HD-VILA-100M dataset as the foundation for its own refined dataset. Panda-70M is a collection of 3.8 million videos from YouTube split into approximately 70.7 million clips and paired with text captions. It was compiled by Snap and released in 2024. Developers used AI to create a new set of captions describing what is pictured in each clip.

93. The HDVILA-100M, HD-VG-130M, Panda-70M, and HowTo100M datasets consist of location identifiers that point to millions of YouTube videos or clips. None of them contains the underlying audiovisual files. To use them in training, a company must access, retrieve and download every referenced video directly from YouTube. Defendant used these datasets to initiate millions of individual downloads of protected YouTube content, all without authorization or compensation, all in violation of YouTube's TPMs, and all for the commercial purpose of building

its foundational video model.

D. Defendant's Automated Scraping Required Circumvention of YouTube's Technological Protection Measures.

94. Bulk extraction of YouTube videos at the scale alleged here cannot occur without circumventing YouTube's layered technological protection measures ("TPMs"), each of which independently controls access to the underlying audiovisual files. Illicit tools and services are specifically designed to defeat these protections, enabling automated systems to retrieve permanent copies of content that YouTube makes available only for streaming. This process is commonly known as "scraping" or "stream ripping." Under 17 U.S.C. § 1201(a), circumvention of any single technological measure that effectively controls access to a copyrighted work is sufficient to establish liability; YouTube's protections are layered and overlapping, and the circumvention of one does not negate the effectiveness or legal significance of the others.

95. "Scraping" or "stream ripping" content involves the automated accessing and downloading of audio and video files directly from a website's servers rather than viewing the content through the provider's authorized playback environment.

96. Upon information and belief, in order to acquire "high-quality text-to-video generation," Defendant, directly and through agents, contractors, and affiliates, intentionally accessed large volumes of YouTube videos by scraping and/or using tools and workflows that bypass or evade YouTube's TPMs and usage restrictions, and then reproduced those videos to assemble training corpora for Defendant's AI models and services.

97. When Defendant scraped audio and video files from YouTube, Defendant did not simply access and download those files onto the YouTube app for offline streaming, as envisioned by YouTube's Premium plan. Instead, Defendant improperly accessed the underlying audio and video files and downloaded those files into Defendant's own system, where Defendant had access over the files and where Defendant could store them indefinitely. These actions circumvented YouTube's TPMs, violated YouTube's Terms of Service, and are inconsistent with the access provided by YouTube's Premium plan.

98. Defendant also obtained its own separate audiovisual datasets directly through its

own independent scraping of YouTube's video sharing platform.

99. To retrieve audiovisual files at scale, Defendant was required to defeat TPM (1), the rolling cipher that protects the true media file URL. On information and belief, Defendant used one or more descrambling tools designed to defeat YouTube's proprietary signature-transformation logic, allowing automated systems to generate valid file requests outside the authorized player environment and thereby obtain the underlying media files directly. Such tools include, for example, the open-source program yt-dlp, which is specifically engineered to descramble, replicate, or extract YouTube's proprietary signature-transformation logic and circumvent the access controls YouTube uses to prevent unauthorized downloading.

100. A descrambling tool such as yt-dlp works by replicating or extracting the signature-transformation process that YouTube uses to protect its media URLs, allowing a user to bypass the authorized streaming environment entirely and retrieve the underlying video and audio files directly from YouTube's servers. Tools of this kind can be used to automatically merge separate audio and video streams and to download entire playlists at scale. On information and belief, Defendant used yt-dlp or a substantially similar descrambling tool for precisely this purpose.

101. To deploy a descrambling tool such as yt-dlp, a user must first obtain and configure the program, along with ancillary software necessary to merge video and audio streams into complete audiovisual files. Once configured, the tool can be directed at individual videos or entire playlists by supplying the Uniform Resource Locator ("URL") for each target video. The result is a complete audiovisual file retrieved outside of and in circumvention of YouTube's authorized playback environment.

102. In simpler terms, a descrambling tool such as yt-dlp acts as a bootleg key for the lock imposed by YouTube's rolling cipher. On information and belief, Defendant used yt-dlp or a substantially similar tool to improperly access and download audiovisual files from YouTube and merge them into complete audiovisual packages for ingestion into their generative AI training pipeline.

103. Defendant also defeated TPM (2), YouTube's IP-based monitoring and blocking system, which restricts high-volume accessing and downloading from a single source. On

information and belief, Defendant implemented an IP-rotation scheme so that when YouTube detected automated activity and blocked one address, accessing and downloading could immediately resume from another. This allowed continuous access to audiovisual files despite YouTube's efforts to detect and terminate abusive activity.

104. YouTube uses automated programs to monitor activity from IP addresses, detect high-volume accessing and downloading behavior, and block those IP addresses from further access to audiovisual content on the platform. An IP-rotation scheme is specifically designed to defeat this monitoring system by cycling through different IP addresses to avoid detection and blocking.

105. Executing an IP-rotation scheme at the scale necessary to download the volume of videos required to train Sora necessarily required the use of virtual machines or substantially similar infrastructure. A single physical machine with a fixed IP address would have been detected and blocked almost immediately given the volume of requests involved. On information and belief, Defendant deployed virtual machines, cloud computing infrastructure, or substantially similar technology to rotate IP addresses at scale, ensuring that when YouTube detected and blocked one address, Defendant's downloading operation could immediately resume from another.

106. On information and belief, by repeatedly cycling through IP addresses in this manner, Defendant ensured that their automated downloading operation could continue uninterrupted even after YouTube detected and blocked individual addresses. The sheer volume of videos necessary to train Sora makes any alternative explanation implausible. The full details of the specific infrastructure Defendant used to execute this scheme will be the subject of discovery.

107. Processes such as these allowed Defendant to bypass YouTube's player page and avoid YouTube's monitoring systems in order to access and scrape content from YouTube.

108. Defendant's conduct necessarily circumvented TPM (3), the use of session-bound, short-lived media URLs. Because these URLs expire and cannot be reused, automated tools must repeatedly obtain fresh authorization parameters and regenerate valid links in order to continue downloading. By programmatically renewing access credentials outside ordinary playback sessions, Defendant maintained uninterrupted access to files that would otherwise become unavailable once the original authorization lapsed.

109. At scale, Defendant's automated activity would also trigger TPM (4), CAPTCHA challenges intended to distinguish human users from automated systems. To continue retrieving videos without interruption, Defendant's infrastructure distributed requests across numerous machines and operated in a manner designed to avoid or neutralize such verification barriers, thereby obtaining audiovisual data without the human interaction required for continued access.

110. Defendant further circumvented TPM (5), YouTube's proof-of-origin token system, which requires requests for video segments to present credentials demonstrating that they originate from an authorized playback environment. Automated downloading tools operate outside that environment and therefore must extract, replicate, or reuse the necessary request parameters in order to retrieve video data directly, allowing Defendant to access and obtain files that would otherwise be delivered only to approved clients.

111. Defendant did not obtain the consent of YouTube, Plaintiffs or other Class Members to conduct its scraping activities.

112. On information and belief, and based on the sheer volume of video data included, the datasets of videos that included YouTube videos were scraped without permission from YouTube and in violation of YouTube's Terms and Services.

113. The automated system used by Defendant to access video content from YouTube was intentionally designed and utilized to circumvent YouTube's TPMs.

114. To be clear, Defendant's generative AI models are not "watching" YouTube in order to be trained. Rather, data improperly accessed and downloaded from YouTube is being uploaded into Defendant's generative AI models to develop and improve Defendant's products.

115. Defendant knew or should have known its conduct contravened YouTube's rules and the TPMs that enforce them.[14]

116. Defendant's actions constitute a breach of the Copyright Act's anti-circumvention provisions, which state, among other things, that "[n]o person shall circumvent a technological measure that effectively controls access to a work protected under this title." 17 U.S.C. §

[14] YouTube's Terms of Service specifically state that users "may access and use the Service as made available to you, *as long as you comply with this Agreement* and applicable law." "Terms of Service," YouTube, https://www.youtube.com/t/terms (last viewed March 12, 2026).

1201(a)(1)(a).

E. Defendant's Actions Caused Harm to Plaintiffs and Class Members

117. At no time did Defendant seek or obtain Plaintiffs' or Class Members' authorization to access and use the videos Defendant used to train its generative AI models.

118. At no time did Defendant seek or obtain Plaintiffs' or Class Members' authorization, nor did Defendant compensate Plaintiffs or the Class Members for the access, copying, ingestion, and use of their YouTube videos in AI training and related workflows.

119. Plaintiffs and Class Members used YouTube as their platform of choice to upload their video content in substantial part due to YouTube's Terms of Service that prohibit the type of scraping, unauthorized accessing and downloading, bulk extraction, and other forms of data mining of audiovisual content utilized by Defendant to obtain Plaintiffs' video content.

120. The harm Defendant is causing goes beyond the immediate economic consequences of unauthorized and uncompensated extraction and use of Plaintiffs' and Class Members' works. It degrades the rights of creators to control their works, determine whether future uses of their work align with their values, and decide the products or services with which they wish to be associated.

CLASS ACTION ALLEGATIONS

121. Plaintiffs bring this action on behalf of themselves and on behalf of all other persons similarly situated.

122. Plaintiffs propose the following Class definition, subject to amendment as appropriate:

All individuals and entities in the United States whose videos were posted on YouTube and were accessed by Defendant.

123. Excluded from the Class are Defendant's officers and directors, and any entity in which Defendant has a controlling interest; and the affiliates, legal representatives, attorneys, successors, heirs, and assigns of Defendant. Excluded also from the Class are Members of the judiciary to whom this case is assigned, their families and Members of their staff.

124. Plaintiffs reserve the right to amend or modify the class definition with greater specificity or division after having an opportunity to conduct discovery. The proposed Class meets

the criteria for certification under Rule 23 of the Federal Rules of Civil Procedure.

125. Plaintiffs also reserve the right to seek certification of subclasses, or alternative classes as shall become necessary in the course of litigation.

126. Numerosity. The Members of the Class are so numerous that joinder of all of them is impracticable. The exact number of Class Members is unknown to Plaintiffs now, but Plaintiffs estimates that there are thousands of Class Members.

127. Commonality. There are questions of law and fact common to the Class, which predominate over any questions affecting only individual Class Members. These common questions of law and fact include, without limitation:

a. Whether Defendant used an automated tool to download or otherwise obtain unauthorized access to the Plaintiffs' and Class Members' copyrighted content on YouTube;

b. Whether Defendant's downloading of the video content was intended to bypass the TPMs put in place by YouTube at the authorization of Plaintiffs and Class Members;

c. Whether Defendant used Plaintiffs' and Class Members' original content for purposes of training its generative AI models;

d. Whether Plaintiffs and Class Members suffered damages from Defendant's misconduct;

e. Whether Plaintiffs and Class Members are entitled to actual damages, statutory damages and/or injunctive relief.

128. These issues are common to the Class, and their resolution would advance this matter and the parties' interests therein.

129. Typicality. Plaintiffs' claims are typical of those of other Class Members because Plaintiffs' video content, like that of every other Class Member, was made available on YouTube and improperly altered and used by Defendant to train Defendant's generative AI models for commercial purposes. Plaintiffs' claims are typical of those of the other Class Members because, among other things, all Class Members were injured through the common misconduct of Defendant. Plaintiffs are advancing the same claims and legal theories on behalf of themselves and all other Class Members, and no defenses are unique to Plaintiffs. Plaintiffs' claims and those of Class

Members arise from the same operative facts and are based on the same legal theories.

130. Adequacy of Representation. Plaintiffs will fairly and adequately represent and protect the interests of the Members of the Class. Plaintiffs' Counsel is competent and experienced in litigating class actions.

131. Predominance. Defendant has engaged in a common course of conduct toward Plaintiffs and Class Members, in that Plaintiffs' and Class Members' video content was unlawfully downloaded, altered and used by Defendant in the same way. The fact that Sora was a foundational model underscores that Defendant's conduct was uniform across the Class: every video unlawfully accessed from YouTube was ingested into the same core model that underpins Defendant's product ecosystem. This commonality further demonstrates the predominance of shared legal and factual issues among Class Members. The common issues arising from Defendant's conduct affecting Class Members set out above predominate over any individualized issues. Adjudication of these common issues in a single action has important and desirable advantages of judicial economy.

132. Superiority. A Class action is superior to other available methods for the fair and efficient adjudication of the controversy. Class treatment of common questions of law and fact is superior to multiple individual actions or piecemeal litigation. Absent a class action, most Class Members would likely find that the cost of litigating their individual claims is prohibitively high and would therefore have no effective remedy. The prosecution of separate actions by individual Class Members would create a risk of inconsistent or varying adjudications with respect to individual Class Members, which would establish incompatible standards of conduct for Defendant. In contrast, the conduct of this action as a class action presents far fewer management difficulties, conserves judicial resources and the parties' resources, and protects the rights of each Class member.

133. Defendant has acted on grounds that apply generally to the Class as a whole, so that class certification, injunctive relief, and corresponding declaratory relief are appropriate on a Class-wide basis.

134. Ascertainability. Finally, all members of the proposed Class are readily ascertainable. Defendant used datasets that contain the full lists of URLs and video identifiers for every YouTube video incorporated into the training pipeline. Those identifiers allow the parties to

determine exactly which videos were used and to match each video to its creator through YouTube's public channel and authorship information. Because the datasets provide a complete map of the videos Defendant downloaded, the identities of the affected creators are identifiable through straightforward reference to the URLs and corresponding channel data.

CLAIM FOR RELIEF

Violation of the DMCA (Anti-Circumvention) 17 U.S.C. § 1201(a)

135. Plaintiffs repeat, reallege, and incorporates the allegations contained in the previous paragraphs as if fully set forth herein.

136. YouTube's Terms of Service prohibit scraping, unauthorized downloading, bulk extraction, or other forms of data mining of audiovisual content. These restrictions operate together with TPMs to prevent unlicensed access to the underlying video files, not simply reproduction of content.

137. Plaintiffs and Class Members uploaded their original video content to YouTube's video sharing platform due in substantial part to YouTube's Terms of Service and TPMs prohibiting bulk downloading of creators' videos.

138. Content creators, including Plaintiffs and Class Members, expect that their works will not be accessed at the file level and copied at scale without consent. Plaintiffs and Class Members own all rights, title, and interest in and to the claims asserted in this action, including all claims for violation of 17 U.S.C. § 1201.

139. Each of Plaintiffs' and Class Members works were uploaded to YouTube. When published, YouTube automatically applied TPMs that control access to and prevent unauthorized copying or downloading of these audiovisual files.

140. YouTube's Terms of Service and associated access controls constitute "effective technological measures" within the meaning of 17 U.S.C. §1201.

141. Defendant used automated tools for the sole purpose of circumventing YouTube's technological barriers that effectively control access to extracted files never made available to the public. Specifically, Defendant employed measures to access, extract, copy, and download Plaintiffs' and Class Members' content from YouTube without authorization. In doing so, Defendant

improperly obtained millions of videos from YouTube's platform.

142. This distinction is critical: viewing a YouTube video through YouTube's authorized and advertising-supported playback mechanisms does not provide access to the underlying file. Defendant's circumvention tools broke through that access barrier, triggering liability under §1201.

143. Plaintiffs and Class Members have been harmed by Defendant's violations of 17 U.S.C. §1201(a) because Defendant has taken their content without authorization or compensation. Defendant's circumvention of YouTube's technological measures has facilitated Defendant's ongoing and mass-scale copyright infringement through its unauthorized and uncompensated use of the content in its training data.

144. Defendant accessed, downloaded, stored and utilized those videos to assemble training corpora for Defendant's AI models and services.

145. By using circumvention software and disabling or bypassing YouTube's TPMs, Defendant violated 17 U.S.C. § 1201(a), which prohibits the circumvention of a technological measure that effectively controls access to a copyrighted work.

146. Each act of circumvention constitutes a separate violation of §1201. Plaintiffs and the Class Members are entitled to statutory damages, injunctive relief, impoundment, and attorneys' fees and costs under 17 U.S.C. §1203.

147. Each of Defendant's acts of infringement is a willful violation as Defendant specifically utilized automated tools designed to evade YouTube's TPMs.

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs respectfully request a judgment in their favor and against Defendant as follows:

A. For a declaration that Defendant has willfully circumvented the copyright protection systems of YouTube intended to protect Plaintiffs' and the Class Members's audiovisual content.

B. For statutory damages (up to the maximum allowed by law per violation), injunctive relief, and attorneys' fees and costs under 17 U.S.C. §1203;

C. For such equitable relief under Title 17, Title 28, and/or the Court's inherent

authority as is necessary to prevent or restrain infringement of Plaintiffs' and the Class Members' copyright-protected content, including a preliminary and permanent injunction requiring that Defendant and its officers, agents, servants, employees, attorneys, directors, successors, assigns, licensees, and all others in active concert or participation with any of them, cease infringing, or causing, aiding, enabling, facilitating, encouraging, promoting, inducing, or materially contributing to or participating in the infringement of any of Plaintiff's or the Class Members' exclusive rights under federal law, including without limitation in the content identified in Exhibits A, B, and C;

D. For an award of pre-judgment and post-judgment interest, to the fullest extent available, on any monetary award made part of the judgment against Defendant; and

E. For such other and further relief as the Court may deem just and proper.

JURY DEMAND

Plaintiffs demand a trial by jury on all claims for which trial by jury is proper.

Dated: April 3, 2026

HEAH BAR-NISSIM LLP

By /s/ ROM BAR-NISSIM
ROM BAR-NISSIM

ELLZEY KHERKHER SANFORD MONTGOMERY LLP

JARRET LEE ELLZEY
TOM KHERKHER
LEIGH S. MONTGOMERY

Attorneys for Plaintiffs