BSA Comments Regarding Intellectual Property Rights in the Era of AI

November 2, 2023

BSA | The Software Alliance (**BSA**) welcomes this opportunity to provide comments to the Intellectual Property Strategy Headquarters (**IP Strategy HQ**) in response to the public consultation on "Intellectual Property Rights in the Era of AI".[1]

BSA is the leading advocate for the global software industry before governments and in the international marketplace[2] and has long supported effective copyright protection.

As the IP Strategy HQ acknowledges[3], AI has the potential to accelerate digitization and the use of digital technology and contribute not only to improving productivity in all of Japan, but also to solving various social issues. The benefits that AI provides encompass all industry sectors, including content creators — consistent with copyright law.

Given the critical importance of AI to long-term economic growth and stability of the country, it is important for the Government of Japan to continue monitoring legal developments to ensure that the Copyright Act and relevant intellectual property laws continue to foster the development of AI.

We provide answers below to questions (1) and (3) listed in the questionnaires from the IP Strategy HQ, to support the Government achieve its goal of enabling digitalization across society to respond to societal challenges and to continue fostering an environment to enable innovation.

(1) What are your views on the relationship between generative AI and copyright?

While AI covers a broad set of technologies and use cases, AI systems learn from the computational analysis of large amounts of data that can be used to train an AI model. The model will then make predictions when presented with a query that includes new data. Some generative AI models can take this a step further and generate new data, such as text or a picture or even new code. Regardless of what form of AI is being used, respect for copyrighted works — as well as personal data privacy and trade secret information — should be a cornerstone of responsible AI development.

[1] https://public-comment.e-gov.go.jp/servlet/Public?CLASSNAME=PCMMSTDETAIL&id=095230820&Mode=0

[2] BSA's members include: Adobe, Alteryx, Atlassian, Autodesk, Bentley Systems, Box, Cisco, CNC/Mastercam, Databricks, DocuSign, Dropbox, Elastic, Graphisoft, IBM, Informatica, Juniper Networks, Kyndryl, MathWorks, Microsoft, Okta, Oracle, Palo Alto Networks, Prokon, PTC, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, TriNet, Twilio, Unity Technologies, Inc., Workday, Zendesk, and Zoom Video Communications, Inc.

[3] https://www.kantei.go.jp/jp/singi/titeki2/ai_kentoukai/gijisidai/siryou3.pdf

22F Shibuya Mark City
1-12-1 Dogenzaka, Shibuyaku,
Tokyo 150-0043

P +81 3 4360 5473
F +81 3 4360 5301
W bsa.org

Japan Representative Office

While there are different approaches to AI training, many large language models (**LLMs**) convert data (e.g., text, images, or other data) from training data into "tokens", and find statistical correlations between this "tokenized"data.[4] The source data may originate from a legally accessed data set, or from content freely available online, or from other sources, and may be subject to copyright protection or not, but the use of such data in training an AI model is unrelated to the expressive content of any particular work. A book may be used to learn language skills, which are then used for improved database management practices. These uses fall as "exploitation without the purpose of enjoying the thoughts or sentiments expressed in a work" as stipulated in Article 30-4 of Copyright Act.[5]

Given the diverse range of AI model development practices and potential use cases, Article 30-4 of Copyright Act remains a doctrine flexible enough to address the highly fact dependent issues around the use of works to train AI machine learning models. Existing legislation, especially Article 30-4 of Copyright Act, provides the correct guidance both in the protection of works as well as the flexibility necessary to develop and produce new and useful services and works in the public interest. We support Japan maintaining this legal system that enables AI developers to use publicly available content to which they have lawful access to train AI systems.

To answer the question on the relationship between generative AI and copyright, we first provide an explanation of the application of AI, the AI development lifecycle, and use of copyrighted works to train AI models to facilitate understanding on this topic. We also follow by providing our views on the copyrightability of works generated using AI systems and potential liability for infringing works generated using AI systems.

A. Artificial Intelligence in a Copyright Context

AI machine learning encompasses a vast array of technologies developed or deployed for use in a variety of different industries and applications. Machine learning depends upon the computational analysis of training data to identify correlations, patterns, or other metadata in order to develop a model that can make predictions or recommendations based on future data inputs.[6] Generative AI models are able to generate new text, image, or sound.

A-1: Applications of AI

To explain the insights, predictions, and other outputs derived from computational analysis in the machine learning context, we provide below a few widely recognized examples:

- Automated flight management and air traffic control based on computational analysis of meteorological conditions, real-time fuel consumption, aircraft operational data, nearby air traffic conditions, airport congestion, and numerous other data elements.[7]

[4]. https://www.axios.com/newsletters/axios-ai-plus-ff6d6e95-f062-495e-aa45-813f95ceae11.html?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axioslogin&stream=top

[5] Article 30-4 of Copyright Act: https://www.japaneselawtranslation.go.jp/ja/laws/view/4207#je_ch2sc3sb5at4

[6] BSA | The Software Alliance, *Confronting Bias: BSA's Framework to Build Trust in AI* (2022), at https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai; *See also,* National Institute of Standards and Technology, *NIST Risk Management Framework* (Jan. 2023), at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (describing an AI system "as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments.")

[7] M. Durgut, *Artificial Intelligence and Air Traffic Control*, Aviationfile.com website (Jan. 2023), at https://www.aviationfile.com/artificial-intelligence-and-air-traffic-control/#:~:text=One%20of%20the%20primary%20applications%20is%20to%20help,make%20informed%20decisions%20on%20routing%20and%20scheduling%20flights; Degas et al., *A Survey on Artificial Intelligence and explainable AI in Air Traffic Management*, 12 Applied Sciences 1295 (2022), at https://www.mdpi.com/2076-3417/12/3/1295; Hanneke Weitering, *How Artificial Intelligence is Transforming Aviation*, Futureflight.aero website (2023) at https://www.futureflight.aero/news-article/2023-07-13/beyond-automation-how-artificial-intelligence-transforming-aviation

- Identification of chemical and cellular anomalies for early diagnosis, prevention, and treatment in the fields of oncology, autoimmune disorders, and Parkinsons and Alzheimers disease.[8]

- By integrating generative AI into security operations, organizations can effectively identify and address security anomalies, as well as detect and mitigate potential threats.

- Predictive climate modeling based on computational analysis of satellite data, weather station data, topographical information, and various IoT and sensor data.[9]

- Improved carbon tracking and mitigation based on computational analysis of transportation logs, meter readings, fuel purchase records, atmospheric pollution tracking, and visual monitoring of power plants and other facilities, and other data sources.[10]

- Computational analysis to map vulnerable seaside areas to produce cyclone risk maps and guide investment plans for cyclone shelters, schools, health facilities, and other infrastructure for disaster planning and survivability.[11]

- Predictive typing and other office productivity solutions (e.g., an "auto-complete" function that suggests the letters "…cerely yours" after the typist inputs the letters, "sin"),[12] or the creation of sound effects or special effects to assist creators and film producers in developing new artworks.[13]

- Generative AI can also enable conversational interfaces to enable people to work with AI across a broad array of domains and applications, such as the applications referred to above. Conversational interfaces can enable people to query and reason over data and outputs of AI systems.

While the use cases are diverse, the elements of training each are very similar as further discussed below. The purpose of training is to enable the model to learn the unprotectable elements of copyright works they are trained on.

A-2: The AI Development Life Cycle

The AI development life cycle typically includes the following steps:

- Project Conception: First, the AI development team will formulate the "problem" that a system is intended to address and map the structure and target variables that the system is intended to

[8] Hunter et al., *The Role of Artificial Intelligence in Early Cancer Diagnosis*, 14(6) Cancers 1524 (2022), at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8946688/; Stafford et al., *A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases*, 3 NPJ - Digital Medicine 30 (2020), at https://www.nature.com/articles/s41746-020-0229-3; Diogo et al., *Early diagnosis of Alzheimer's disease using machine learning*, 14 Alzheimers Research and Theory 107 (2022), at https://alzres.biomedcentral.com/articles/10.1186/s13195-022-01047-y;

[9] Schneider et al., *Harnessing AI and computing to advance climate modelling and prediction*, 13 Nature Climate Change 887 (2023), at https://www.nature.com/articles/s41558-023-01769-3; World Economic Forum, *The role of machine learning in helping to save the planet* (2021), at https://www.weforum.org/agenda/2021/08/how-is-machine-learning-helping-us-to-create-more-sophisticated-climate-change-models/

[10] Global Data Alliance, *Cross-Border Data Transfers & Environmental Sustainability* (2023) (internal citations omitted), at https://globaldataalliance.org/wp-content/uploads/2023/04/04192023gdacbdtsustainability.pdf

[11] Global Data Alliance, *Cross-Border Data Transfers & Environmental Sustainability* (2023)

[12] S. Ashraf, *Desmystifying Autocomplete,* Towards Data Science (2020), at https://towardsdatascience.com/index-48563e4c1572; A. Wickramarachchi, *Machine Learning: Word Embedding and Predicting,* Towards Data Science (2020), at https://towardsdatascience.com/machine-learning-word-embedding-and-predicting-e603254e4d7b

[13] *See e.g.,* D. Nelson, *AI Researchers Design Program To Generate Sound Effects For Movies and Other Media*, Unite AI Website (2022), at https://www.unite.ai/ai-researchers-design-program-to-generate-sound-effects-for-movies-and-other-media/

predict. For models trained for a particular task, sometimes referred to as narrow AI applications, this may be a fitness app that analyzes a consumer's heart rate to monitor for irregularities that might predict whether that person is at risk of a stroke or heart disease (i.e., the target variable). Foundation models will however be developed to power a broad array of applications.

- Raw Data Identification: Second, the AI development team may identify a relevant universe of "raw data" that will be subsequently transformed and structured. Data sources are as diverse as the potential applications of machine learning AI and may include everything from machine-to-machine data (e.g., satellite transmission data) and international trade statistics to published materials, blog posts, website comments, and chat room logs. "Raw data" is frequently "messy," requiring significant work to transform the data into a usable form, as outlined below. The data to develop a narrow AI system for a specific application or fine-tuning a foundation model for a particular task, will relate to the particular task. Foundation model training will require broader and more varied data. The scale of data needed to train large language models is vast and will rely on being able to analyze data on the internet to achieve this scale.

- Preparing the Data Set: The AI development team may modify the "raw data" so that it can be understood semantically by the machine and used to train the model. During this process, the team will revise, clean, and normalize the data as necessary. Data typically is transformed semantically and structurally through "tokenization," which involves breaking down a piece of text or data into smaller units (or "tokens") for purposes of computational analysis. Additional processing may be necessary to improve the reliability, quality, and suitability of the data for analysis, helping to address quality challenges such as missing values, duplicates, outliers, and inconsistent formatting across the entire data set. Large scale AI models are typically trained using self supervised methods, dispensing with the need to label the training data and allowing the models to scale on a vast scale of training data, typically collected from the internet. Training data often will appear ultimately as a sequence of tokens that stretches across the entire AI corpus.

- Model definition: After input data has been suitably processed, the AI development team must establish the system's underlying architecture. This includes identifying the variables (i.e., "features") in the training data that the algorithm will evaluate as it looks for patterns and relationships as the basis of a rule for making future predictions. It also includes selecting the type of algorithmic model that will power the system (e.g., linear regression, logistic regression, deep neural network.)[14] Once the data is ready and the algorithm is selected, the team will train the system to produce a functional model that can make predictions about future data inputs.

- Model Validation, Testing, and Revision. After the model has been trained, the AI development team must validate it to determine if it is operating as intended and test it to demonstrate that the system's outputs fall within expected parameters and do not contain unexpected errors or unintended bias. Based on the outcome of validation and testing, the team may need to revise and refine the model to mitigate these risks.

B. The Use of Copyrighted Works to Train AI Models

As discussed above, computational analysis is typically applied to a large training data corpus that may comprise millions or billions of tokenized data elements. Depending on how the model is trained, data accessible over the Internet may be collected as part of the raw data set that is transformed into that corpus.

This raw data may include copyrighted works because a substantial portion of the content freely available on the Internet is potentially subject to copyright protection, which has a low threshold to establish "originality" and which provides that copyright arises automatically upon a work's creation, even if it the

[14] BSA AI Bias Framework, *infra* note 6.

work is not registered. Importantly, however, not all the material online is subject to copyright, in part because copyright protection does not extend to facts, ideas, or mathematical concepts.[15]

Computational analysis may involve two sets of reproductions that potentially implicate the Copyright Act: (1) reproductions necessary to create a corpus of "training data," and (2) transient reproductions that are incidental to the computational process of training the AI model. In each case, the reproductions are "intermediate" in the sense that they are not visible or otherwise made available to the public. Instead, the reproductions are the necessary byproduct of a technical process that is aimed at identifying non-copyrightable information *about* the underlying corpus of works — i.e., the correlations and patterns that inform the creation of the AI model and enable it to make predictions based on future data inputs. Such intermediate, non-expressive reproductions do not compete with and have no impact on the economic interests that copyright is intended to protect.

Furthermore, computational analysis does not involve the consumption of any copyrighted works for their expressive content. Rather, such analysis involves mathematical calculations of probabilities, correlations, trends, and other patterns across the entire tokenized data set. Such analysis seeks to understand only the mathematical patterns (e.g., the relationships of specific tokens in relation to other tokens) distributed across the entire data set. These mathematical patterns are themselves not expressive content protected by copyright law.

As noted below, established legal doctrine ensures that such reproductions are generally permissible.

B-1: Computational Analysis of AI Training Data is a Data Analysis

The exception provided in Article 30-4 ensures the use of copyrighted works for the purposes of analyzing a large collection of information to identify patterns, correlations, and other metadata to develop an AI model that makes predictions about future data inputs.

By way of illustration, this means that an AI developer seeking to create a natural language processing model — such as an AI-driven predictive typing model — can rely on publicly available text-based material to create the training database. In such a scenario, the AI developer would not be reproducing this text for its expressivepurpose. Rather, the reproductions would be made solely for the purpose of extracting unprotected information about the correlations, patterns, and relationships among letters and words as they appear in thousands of phrases, figures of speech, similes, metaphors, grammatical patterns, and common linguistic formulations and expressions. Neither the letters, words, and phrases, nor the mathematical patterns among them across thousands or millions of writings, are copyright protectable subject matter.

Similarly, for an image generation model, the model developer will take a very large volume of images tagged with words — for instance, some number of cat photos tagged with "cat," and dog photos tagged with "dog." Over time, the model learns that certain patterns are characteristic of "cattiness" and it will learn to recognize whether an image fed to it is a cat or not, whether the image is a cartoon, or a photograph, etc. Once again, the machine is not reproducing for expressive use any particular image of cats but is instead dissecting the images to understand what a cat is — a basic set of facts/statistical correlations rather than expressions.

Furthermore, the ultimate use of the computational analysis applied to the data set is a transformative use of the original content. Auto-complete functionality in predictive typing software comprises a new creation in the form of software code that is transformatively distinct both from the entire AI training corpus and from any single work within that corpus. Software code that can suggest the endings of commonly used phrases is many steps removed from any copyrighted works found in the underlying raw data. Finally, such functionality simply does not compete with any copyrighted works in any manner that copyright is

[15] Article 9.2 of the TRIPS Agreement provides that copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such. In other words, copyright protection does not cover any information or ideas contained in a work; it only protects the original way in which such information or ideas have been expressed. Thus, everyone is free to use the information contained in a work, including for the purpose of creating new works.

intended to protect. Accordingly, the AI developer can safely rely on the copyright exception to construct an AI training database using text, images, or other data to which he/she has lawful access.

C. The Copyrightability of Material Generated by an AI System

With regards to the copyrightability of works generated using AI systems, the analytical touchstone should be whether human creativity was responsible for the work regardless of what instrument or technology was used to aid its expression. Generative AI should bolster creativity, just as other software applications have long been an important tool of artists and storytellers. Generative AI is used, for example, in word processing for authors and photo enhancement for visual artists; it is used to create special effects in audio-visual works and arranging music for sound recordings. When generative AI is used to enhance human creativity, the resulting work should be protected by copyright.

As inquiries regarding copyrightability will turn on a close examination of the degree of human creativity, and with artificial intelligence increasingly being used as a tool in all categories of creative works, a decision to limit copyrightability when AI is used would significantly chill adoption of AI solutions.

Works that emerge as outputs of AI systems and meet the human creativity requirement should continue to be eligible for copyright protection. In most cases, AI systems will function as tools used by human authors and creators to execute upon their creative vision. For instance, photographers will use AI-enabled tools to automate the tedious process of editing their images,[16] architects will use AI to augment their designs to enhance their energy efficiency,[17] and filmmakers will use AI to ensure that the movement of their animated characters appear more life-like.[18] In each of these cases the creative contribution of the human user makes it easy to conclude that the output would be copyrightable.

The use of generative AI should not change the analysis. Certainly, there will be extreme cases in which it is either clear that there is no spark of human creativity involved or, on the other hand, generative AI was not part of the creative expression.

D. Potential Liability for Infringing Works Generated Using AI Systems

In relation to the potential infringement of works generated using AI systems, BSA offers the following observations.

D-1: Current Copyright Law Protects Copyright Holders from Infringement, Including in Cases Arising from AI Generated Content

Copyright holders should have full and effective remedies when their rights are infringed. This principle applies equally to outputs generated using AI systems and outputs generated in other ways. Whenever such infringement is found, it is critical to fully compensate artists and creators for any damages caused.

In our view, existing copyright law should prove adequate to address questions of infringement. In most AI use cases, such as those described above, the output of an AI system will not implicate copyright at all. However, AI, like other technologies, could be used to create infringing material. In those situations, infringement liability would be premised upon proof that the allegedly infringing output is based on "actual copying" and is "substantially similar" to the copyrighted work, such that an "ordinary reasonable person would fail to differentiate between the two works" in view of the "qualitative[ly] and quantitative[ly] significant" similarity between the works. In cases in which the copyright claims are "thin," because there is only one way, or there are only a few ways, to represent specific facts, it would be necessary to prove "virtual identity" between the output and copyrighted work, consistent with prevailing legal norms. Many users of generative AI will be small businesses experimenting with new tools, and enforcement decisions

[16] https://theblog.adobe.com/adobes-general-counsel-makes-the-case-for-ai/

[17] https://www.autodesk.com/redshift/machine-learning-in-architecture/

[18] https://theblog.adobe.com/state-of-ai-in-animation/

should take account of this. Liability would also arise in other appropriate cases, such as those involving derivative works.

Plaintiffs may also seek to bring infringement actions against providers of AI-related services. Here too, the existing Copyright Act should prove adequate in evaluating and apportioning liability. If a plaintiff demonstrates that a direct infringement has occurred, courts will evaluate whether the service provider should be deemed "contributorily" and/or "vicariously" liable for its users' conduct.

In the cases where an AI model's output is substantially similar to a work the model has been trained on, the courts have tools at their disposal to make case-by-case determinations as to whether the output is infringing. We believe that existing law sets adequate and fair boundaries and is not cause for policy changes at this time.

D-2: AI Models Trained on Sufficiently Large Data Sets are Less likely to Produce Infringing Outputs

While it is important not to conflate training data with the output of an AI system, it is worth noting that the more data available for training, the less likely the system will produce a copy or derivative of any particular input (in the absence of a user's intent to infringe). In any well-designed AI model trained on a sufficiently large data set, computational analysis should never (or only in the rarest of circumstances) produce outputs that are "substantially similar," let alone "virtually identical," to any specific copyrighted work.

Some AI developers and deployers are already taking steps to engage with artists and creators on how to support their work in a changing digital environment and taking steps to limit misuse of an AI system, such as limiting what prompts can be used.

(3) What are your views on how technology should be handled from the viewpoint of avoiding intellectual property risks related to generative AI?

First, we would like to address that the risk that AI systems will be used to create infringing works, the Copyright Act is sufficiently flexible to determine whether copyright infringement liability exists. Stated differently, whether the output is infringing is related not to the technology used. Where an output is infringing, there should clearly be liability regardless of whether AI is used in producing the infringing copy.

We also want to highlight that there have been calls to create regulatory mechanisms that would restrict in practice access to works online for any purpose, including training an AI system. Automated "do not crawl" tools already exist and can prevent access to works online for any purpose, including training an AI system. We support further voluntary conversations between creators and AI developers and deployers to arrive at effective, consensus technical mechanisms.

Conclusion

BSA appreciates the opportunity to provide comments to IP Strategy HQ. As discussed above, the non-consumptive computational analysis of content in AI training data sets falls under the exception provided in the Copyright Act. At the same time, BSA supports multi-stakeholder efforts relating to AI training processes as well as efforts to minimize the risk of infringement. And to the extent that infringement occurs, BSA strongly supports fully protecting content creators. Finally, BSA supports efforts to explore how the current Copyright Act could better protect against improper digital replication of a person's name, image, likeness, or voice in a manner that competes with his or her professional and commercial interests.

22F Shibuya Mark City
1-12-1 Dogenzaka, Shibuyaku,
Tokyo 150-0043

P +81 3 4360 5473
F +81 3 4360 5301
W bsa.org

Japan Representative Office