BSA Comments on Draft Approach to AI and Copyright

February 12, 2024

General Comments

BSA | The Software Alliance (**BSA**)[1] welcomes this opportunity to provide comments to the Agency for Cultural Affairs (**Agency**) in response to the public consultation on the draft "Approach to AI and Copyright" (**Draft**).[2]

BSA is the leading advocate for the global software industry before governments and in the international marketplace. Our members are among the world's most innovative companies, creating enterprise software solutions including artificial intelligence (AI) and machine learning systems that help businesses of all sizes in every part of the economy to modernize and grow.[3] Representing enterprise software developers, BSA has long supported effective copyright protection.

Artificial intelligence (**AI**) has the potential to accelerate digitization and the use of digital technology. This will contribute to improving productivity in Japan and solving social issues. AI is advancing innovation and creativity in every sector. For example, AI provides creators with new tools to enhance their craft — in special effects in film, in sound mixing, in architectural planning, and in vehicular styling and design. As this technology continues to evolve, it is important to consider the role of copyright law in encouraging innovation and protecting the rights of creators.[4]

Given the critical importance of AI to long-term economic growth and stability in Japan, it is important for the Government of Japan to maximize these benefits and continue to ensure that the Copyright Act (**Act**) and relevant intellectual property laws continue to foster the development of AI and protect copyright owners from infringement.

In this respect, we commend the current efforts of the Agency to provide guidance on the interpretation of the existing Act to respond to the concerns raised by various stakeholders

[1] BSA's members include: Adobe, Alteryx, Altium, Amazon Web Services, Asana, Atlassian, Autodesk, Bentley Systems, Box, Cisco, Cloudflare, CNC/Mastercam, Dassault, Databricks, DocuSign, Dropbox, Elastic, Graphisoft, Hubspot, IBM, Informatica, Kyndryl, MathWorks, Microsoft, Nikon, Okta, Oracle, PagerDuty, Palo Alto Networks, Prokon, Rockwell, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, TriNet, Twilio, Workday, Zendesk, and Zoom Video Communications, Inc.

[2] *Solicitation of Opinions on "Approach to AI and Copyright (Draft)"*, January 23, 2024, at https://public-comment.e-gov.go.jp/servlet/Public?CLASSNAME=PCMMSTDETAIL&id=185001345&Mode=0

[3] *How Enterprise Software Empowers Businesses in a Data-Driven Economy*, January 19, 2021, at https://www.bsa.org/files/policy-filings/011921bsaenterprisesoftware101.pdf

[4] *Artificial Intelligence & Copyright Policy: Advancing Technology and Creativity in the 21st Century Economy*, January 8, 2024, at https://www.bsa.org/files/policy-filings/01082024bsaaicopyright.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

including copyright owners (creators, performers), business operators (AI developers, AI service providers), and AI users. As the Agency acknowledges, the interpretation of copyright law, not only with regards to generative AI, is based on judicial decisions made in each specific case. However, anticipating developments and uses of AI is critical to further accelerate AI adoption in Japan. We provide our comments below to support Japan achieving its goal of enabling digitalization across society including through AI development and adoption to respond to societal challenges and to continue fostering an environment that enables innovation.

We are concerned that the Agency may be conflating inappropriately the output from generative AI systems and the input in the form of training data in many of the sections of the Draft. Large language models (**LLMs**) and other generative AI systems typically are trained with millions if not billions of pieces of data, some of which may be part of a copyrighted work, but they do not store expressive content for retrieval or reproduction. Instead, training data is analyzed for statistical patterns after the content (e.g., raw data) has been converted to into structures that allow for semantic and structural analysis (a process called "tokenization").[5] What is extracted, or "learned" is not the content per se, but the patterns inherent in the content analyzed across the entire training data set of millions or billions of works.

Existing Copyright Act

[2. Premise of Consideration 3. Regarding Technical Background of Generative AI 4. Various Concerns Voiced by Stakeholders]

As stated in the Draft,[6] the current Copyright Act, including the relevant exceptions to exploitation of copyrighted works related to computational data analysis and for training AI models (Articles 30-4, 47-4, and 47-5**)** provides well-balanced guidance both for the protection of copyrighted works and the flexibility necessary to develop and produce new and useful services and works in the public interest.

As stated above, when training AI models, computational analysis typically involves turning data into tokens that are then analyzed for statistical correlations with other tokenized data across the training data set, which may contain millions if not billions of pieces of data. Some of this data may be extracted from copyrighted works, but the use usually has nothing to do with the expressive content of any particular work. In cases where an AI model's output is substantially similar to a copyrighted work, courts have tools at their disposal to make case-by-case determinations as to whether the output is infringing. As such, there is no need to amend the Act at this time.

Training/Development Phase

Regarding Cases That Are Not for the Purpose of Enjoying a Work

[5. Each Discussion Point (1) B. Regarding Cases in Which the Purpose of Enjoyment Coexists with the "Case of Exploitation for Data Analysis"]

Avoid Over-Restriction of the Existing Exceptions for AI Training: While we appreciate the Agency's effort to interpret the Act in the context of AI, including generative AI, we also encourage

[5] *BSA Comments Regarding Intellectual Property Rights in the Era of AI*, November 2, 2023, at https://www.bsa.org/files/policy-filings/11022023iprightsai.pdf

[6] *Draft Section 2.(1) "Premise of Consideration"*

the Agency to avoid overly narrowing or restricting the scope of the existing exceptions for the use of copyrighted works for training AI models under Article 30-4 of the Act.

For example, the Draft states that "the fact that the generation of outputs having common creative expression as the copyrighted work used in training occurs extremely frequently during the generation/utilization phase may become a factor in inferring the existence of the purpose of enjoyment during the development/training phase".[7] This statement indicates that the Agency views the frequency of potential copyright infringing output as directly related to the purpose of use of the training data and that the limitation of rights under Article 30-4 of the Act would not apply.

However, in the generation/utilization phase, it is important to understand that AI users are able to freely input prompts and generate outputs within the scope of the prescribed use. The user's interaction with the model usually is not managed or controlled by the entity that developed or trained the AI model. In most LLMs, no particular output would be "frequent" absent prompts from the user directing the system to produce particular output.

It may be possible for users to trick (e.g., direct) the model by using unusual prompts to generate content that is similar to its training data or other copyrighted works. If an AI model generates outputs substantially similar to a copyrighted work through such prompts, that directed output is where the infringement may occur. Purposeful prompt engineering by a user to solicit infringing outputs should not have a bearing on the validity of using copyrighted works for AI training pursuant to Article 30-4 and other relevant exceptions.

Also, it is important to understand that foundation models support many different applications. A foundation model that is trained by analyzing a large training dataset may be developed by one party and provided to another party that will develop an application that uses the foundation model. The nature of the output that is generated by the eventual use of the model should not be used as a presumption of whether the training data may be considered infringing.

We urge the Agency to refrain from generalizing the relevance of the generation/utilization phase to the evaluation of the Act with respect to the development/training phase. Such an interpretation of the Act would greatly undermine the predictability and legal stability for business operators that develop and train AI and, as a result, would discourage the development and training of AI models in Japan and decrease in investment in AI development and deployment.

Retrieval-Augmented Generation (RAG)

[C. Regarding Retrieval-augmented Generation (RAG), etc.]

The Draft describes an AI methodology referred to as "retrieval-augmented generation" (**RAG**), in which a generative AI model searches for target data, including possibly copyrighted works, and generates a summary of results.[8] The Draft suggests that in certain circumstances, if the output of the RAG includes all or part of the creative expression of data that was "retrieved" for training, then it may be deemed that the initial collection of the data was "for the purpose of enjoyment" of the copyrighted work. In such a circumstance, according to the Draft, Article 30-4 may not apply to the collection and use of such copyrighted works for the purposes of AI training.

Again, on this point, we urge the Agency to avoid conflating training data with the output of an AI system in the Draft. If an AI system generates copyright infringing output, then remedies to such infringement should apply to that output. But this should have no bearing on the use of such

[7] *Draft Section 5.(1)B(b) "Regarding Cases in Which the Purpose of Non-enjoyment and Purpose of Enjoyment Coexist"*

[8] *Draft Section 5.(1)C "Regarding Retrieval-augmented Generation (RAG), etc."*

copyrighted material for training, as the use of that material is not for "enjoyment" of the expressive content, but to extract statistical data regarding the data, which is not subject to copyright. As such, any assessment of copyright concerns regarding RAG should focus on whether the output infringes the copyright of a copyrighted work. The technical processing applied to works either used for training or used as reference for the model are for the purposes of data analysis and would not be copyright infringements. The additional exception under 47-5 (Minor Exploitation Incidental to Computerized Data Processing and the Provision of the Results Thereof), should also be applicable to the output generated for RAG, when the conditions are met. If copyright is found to be infringed by the generated output, this should not have bearing on whether the reproduction made for the purposes of data analysis is infringing. This is true for both the training data and any data that is converted into tabular form for reference.

Commercial Databases

[(c) Examples of Copyrighted Works of Database Organized in a Form That Can be Used for Data Analysis]

The Draft implies that the limitation of rights under Article 30-4 may not apply to the content made available via a commercial database when the rights owner offers a license to enable data analysis through databases.[9] The Agency should clearly state that the Act does not intend to enable a copyright owner to override a copyright exception merely by offering a license. The use of an individual copyrighted work used for AI training does not necessarily undermine the potential market for a commercial database, even if the database is comprised of a collection of copyrighted works. Copyright protection does not extend to data analysis of a copyrighted work if the purpose is not for the enjoyment of creative expression, whether or not the copyrighted work is included in a commercial database.

The Draft further explains that even if measures to block all crawlers for AI training are not taken with robots.txt file, if some measures are taken to block a certain AI training crawler, this will be one of the factors on which it is presumed that a database work, which includes the data in the website and organized in a form that can be utilized for data analysis, is planned to be sold in the future.[10] Given that the Draft presents no court precedents or investigative results that adopt empirical rules or evidence as grounds for this presumption, it is not appropriate for the Agency to provide guidance that could in fact become the grounds for this presumption. The intention of selling databases in the future should be specifically argued and proven by the author of the said database.

Furthermore, in many cases, crawlers are blocked merely from the perspective of ensuring security, protecting data or content, etc. It is difficult to assess the possibility of potential future sales channel being obstructed or to foresee whether AI outputs would share the common creative expression of the copyrighted works which were in the training dataset. As such, we urge the

[9] *Draft Section 5.(1) D.(c) "Examples of Copyrighted Works of Database Organized in a Form That Can be Used for Data Analysis"*

[10] *Draft Section 5.(1)D(d) footnote 24*

Agency to avoid the above broad interpretation evaluating the application of the proviso to Article 30-4 of the Act from these foregoing perspectives, as it could hinder the future of AI development.

Demand for Destruction of a Trained Model

[Regarding Measures Against Infringement/ F. (b) Regarding Demand for Destruction of Trained Model]

The Draft describes that, in certain circumstances, a demand for the destruction of the trained model may be admitted if the model is deemed to be an "object that gives rise to an act of infringement", an "object made through an act of infringement", or a "machine or tool used solely for an act of infringement". As implied in the Draft,[11] considerable amounts of time and money are invested in developing an AI model. Furthermore, an AI model is neither an "object that gives rise", nor is it an "object made" through an act of infringement, and it is not a "machine or tool used solely for" an act of infringement. Suggesting that the destruction of an AI model is an appropriate remedy is not reasonable. The focus of an injunction should be on enjoining the use of an AI system or the act of further creating infringing outputs and using such outputs in an infringing manner, not on the actual model which does not infringe. An AI model is not a collection of reproductions of copyrighted works.

Also, under Japanese law, injunctions generally are allowed only in situations that cannot subsequently be remedied by monetary compensation. The statements in footnotes 29 and 31 of the Draft are therefore misleading as they could be read to indicate that a destruction order may be rendered immediately "in situations in which it is highly possible to generate outputs that are similar to copyrighted works that are used as training data," or "in situations in which an enjoyment purpose and a non-enjoyment purpose both exist" in RAG, etc.[12] Instead, it should be clear that a remedy such as a destruction order would be available only in the extreme and unlikely situation where a plaintiff will suffer irreparable damage and there are no other technological means to mitigate the risk that a model will continue to infringe a work that is the focus of the rights owner's suit.

Generation/Utilization Phase

Determination of Reliance

[(2) B. (b) 2) Cases in Which AI User Was Not Aware of The Existing Copyrighted Work, But the AI Training Data Includes the Said Copyrighted Work]

The Draft states that "if the AI user was not aware of an existing copyrighted work (its expressive content), but the generative AI had been trained on the said copyrighted work during the development/training phase, it can objectively be recognized that there had been access to the said copyrighted work, and therefore, if an output similar to the said copyrighted work is generated by using the generative AI, it is usually considered that reliance is recognized and copyright infringement may be established".[13] The Agency should not make such a general presumption, as the AI system may have created the output based on inferences from non-copyrightable information about the copyrighted work, rather than the work itself. Further, while the cases presented in Draft Section 5 (2)B(b) can assist in determining whether AI generated content may

[11] *Draft Section 5.(1)F.(b) "Regarding Demand for Destruction of Trained Model", footnote 31*

[12] *Draft Section 5.(1)F(b) "Regarding Demand for Destruction of Trained Model", footnotes 29 and 31*

[13] *Draft Section 5.(2)B(b)2) "Cases in Which AI User Was Not Aware of The Existing Copyrighted Work, But the AI Training Data Includes the Said Copyrighted Work"*

be considered copyright infringing, typically, the work used for AI training will be parameterized, and it would be very rare, and therefore difficult, if not impossible, to determine whether a particular generated output relied on one specific work out of billions in the training data. Thus, the Agency should avoid taking for granted "reliance" on training data when assessing whether generated output may be copyright infringing.

Responsible Entity for an Infringing Act

[G. Responsible Entity for an Infringing Act]

The Draft discusses the allocation of responsibilities when AI outputs allegedly infringe others' copyrighted materials.[14] It also proposes some actions for AI developers and deployers to prevent copyright infringements, including the introduction of technological measures to avoid generation of outputs similar to existing copyrighted works.[15] The Agency should avoid recommending liability rules that are not based on current copyright law. Such an approach would also disincentivize developers from releasing foundational AI tools and models, particularly on an open-source basis, and would make AI research highly costly. We do encourage rights owners, artists, and AI deployers to discuss voluntary best practices that could limit infringement.

Other Discussion Points

Disclosure of Training Data

[J. Cases in Which the Disclosure of Copyrighted Works, etc. Used for Training is Required]

The Draft presents cases in which the disclosure of training data may be required to evaluate the reliance of AI outputs on existing copyrighted works.[16] Requiring disclosures could lead to the disclosure of trade secrets or confidential or proprietary information about the design or use of an AI system and therefore discourage investment in AI technology. Forcing AI companies to disclose the contents of these datasets would, in effect, force the publication of valuable and otherwise confidential commercial secrets. There may be appropriate situations in which an AI developer would disclose a summary of the sources of training data, but a detailed accounting would be impractical and should not be required. We recommend the Government instead to focus on advancing the competitiveness of Japan's AI and creative industries by maintaining the core provisions of copyright law that offer technology-neutral protections to legitimate rights ownrs and innovators.

Copyrightability of Outputs

[(3) Copyrightability of Outputs]

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 bolsters 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; in software development, it is used to assist in generating software code based on the programmer's instructions. When generative AI is used to enhance human creativity, the resulting work should be

[14] *Draft Section 5.(2)G "Responsible Entity of Infringing Act"*

[15] *Draft Section 5.(2)G-4) "Responsible Entity of Infringing Act"*

[16] *Draft Section 5.(2)J "Cases in Which the Disclosure of Copyrighted Works, etc. Used for Training is Required"*

protected by copyright. In instances where AI-generated works do not contain creative elements, but are combined with human-authored works, it should not render the entire combined work unprotectable. Instead, otherwise unprotectable AI-generated portions may be disclaimed, but protectable portions of the work should be copyrightable. 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, architects will use AI to augment their designs to enhance their energy efficiency,and filmmakers will use AI to ensure that the movement of their animated characters appear more life-like. 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.

Other Discussion Points

Returning Compensation to Copyright Owners

[(4) Other Discussion Points]

The Draft presents ideas for the means of returning compensation to copyright owners whose copyrighted works are used for AI training.[17] We agree with the view presented in the Draft that it would be difficult to theoretically explain the introduction of a levy system under the Act as a means of returning compensation, as the interests of copyright owners are not usually prejudiced by the exploitation of copyrighted works for data analysis for AI training and development. The “Remaining Issues (for Discussion)”[18] in the Review Committee for Intellectual Property Rights in the Era of AI (5th meeting) states that “It cannot be said that it is appropriate to take uniform measures to return profits; it may be appropriate for the Review Committee to indicate examples of possible voluntary measures that may be discussed by the private sector.” The Agency should take the same approach in the Draft.

Conclusion

BSA appreciates the opportunity to provide comments to the Agency. BSA encourages multi-stakeholder engagement to improve understanding of AI training processes to support efforts to minimize the risk of copyright infringement. To the extent that copyright 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,

[17] *Draft Section 5.(4)*

18 Remaining Issues (for Discussion), Review *Committee for Intellectual Property Rights in the Era of AI at* https://www.kantei.go.jp/jp/singi/titeki2/ai_kentoukai/gijisidai/dai5/siryou1.pdf

image, likeness, or voice in a manner that competes with his or her professional and commercial interests.

In the future, we hope that the Agency will provide sufficient time for consultation. We are disappointed that we and other stakeholders have not been afforded a more reasonable period to translate, study, and consult on this important document. Published on January 23 and with comments due February 12, the consultation has left stakeholders less than three weeks to respond. Such a short time to respond makes it very difficult to provide thoughtful and constructive feedback within the allotted time. We would appreciate at least four weeks and ideally more for detailed documents such as this, for stakeholders to provide substantial, thoughtful, and constructive input.

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