

NUMÉRO 18
DÉCEMBRE 2024

REVUE FRANCOPHONE DE LA **PROPRIÉTÉ INTELLECTUELLE**



ASSOCIATION FRANCOPHONE DE
LA PROPRIÉTÉ INTELLECTUELLE

www.revue-rfpi.com

ISSN 2490-8347

IA et fouille de textes et de données : retour sur la souplesse de l'exception japonaise à l'ère de l'IA générative

AI and Text and Data Mining: revisiting the flexibility of the Japanese exception in the era of generative AI

Elodie MIGLIORE

Doctorant au CEIPI
Université de Strasbourg

Les exceptions de fouille de textes et de données sont souvent évoquées dans les discussions juridiques autour des systèmes d'intelligence artificielle. Si les exceptions prévues par le droit européen sont souvent critiquées, elles sont également comparées aux exceptions d'autres pays. Cet article vise ainsi à préciser l'exception de fouille de textes et de données japonaise dans le contexte de l'IA générative, revenant notamment sur son apparente souplesse.

Text and data mining exceptions are often mentioned in legal discussions dealing with AI systems. The European Union recently created two TDM exceptions. While these exceptions are often criticised, they are also compared to TDM exceptions of other countries. This article aims to analyse the Japanese TDM exception, especially in the context of generative AI and by challenging its apparent flexibility..

Introduction

Les systèmes d'intelligence artificielle (IA) ont créé ces dernières années de nombreux débats juridiques, la propriété intellectuelle n'échappant à la règle¹. L'Europe en a par ailleurs fait une priorité, comme en témoigne de nombreux travaux² et la récente publication au Journal Officiel du très attendu règlement sur l'intelligence artificielle (ci-après RIA)³. Si ce règlement apparaît comme le texte le plus poussé en la

matière, d'autres textes ont d'ores et déjà été évoqué dans les débats relatifs à l'IA.

En effet, les opérations de fouilles de textes et de données sont sous le feu des projecteurs depuis quelques années et, notamment, dans leur relation aux systèmes d'IA. Les articles 3 et 4 de la directive sur le droit d'auteur et les droits voisins dans le marché unique

¹ V. notamment à ce sujet les nombreuses procédures en cours aux États-Unis, par exemple *Getty Images (US), Inc. c. Stability AI, Inc.* (1:23-cv-00135, District Court, D. Delaware).

² V. à ce sujet, Commission européenne, L'intelligence artificielle pour l'Europe, COM(2018) 237 final, Bruxelles, 25 avril 2018 ou encore, Commission européenne, Livre blanc, Intelligence artificielle, Une approche européenne axée sur

l'excellence et la confiance, COM(2020) 65 final, Bruxelles, 19 février 2020.

³ Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle).

numérique⁴ (ci-après DAMUN) ont été – et sont encore – le fruit de nombreuses discussions⁵ et, tout particulièrement, dans leur rapport avec l'intelligence artificielle. Ces articles introduisent deux exceptions de fouille de textes et de données⁶, une étape technique importante et nécessaire pour le développement de systèmes d'IA. Ces articles sont souvent évoqués comme pouvant permettre aux développeurs de systèmes d'IA d'utiliser certaines données pour le développement desdits systèmes, sous réserves de remplir les conditions afférentes.

Cette même directive définit, en son article 2, le processus de fouille de textes et de données comme « toute technique d'analyse automatisée visant à analyser des textes et des données sous une forme numérique afin d'en dégager des informations, ce qui comprend, à titre non exhaustif, des constantes, des tendances et des corrélations »⁷. Cette définition est également complétée par le considérant 8, précisant que la fouille de textes et de données permet une « analyse informatique automatisée d'informations sous forme numérique, telles que du texte, des sons, des images ou des données [...] »⁸. La définition de la directive n'a pas suscité que l'engouement en doctrine et été critiquée à plusieurs égards⁹. D'autres définitions peuvent venir éclaircir ce qu'est un processus de fouilles de textes et de

données comme des définitions plus techniques. La fouille de textes et données peut être définie comme une technique de recherche permettant de collecter des informations à partir de grandes quantités de données numériques grâce à des outils logiciels automatisés¹⁰.

Le texte européen a donc consacré une fouille de textes et de données dite « scientifique », en son article 3. Cet article prévoit une exception pour « les reproductions et les extractions effectuées par des organismes de recherche et des institutions du patrimoine culturel, en vue de procéder, à des fins de recherche scientifique, à une fouille de textes et de données sur des œuvres ou autres objets protégés auxquels ils ont accès de manière licite »¹¹. Cette exception est limitée dans ses bénéficiaires, puisque seuls des organismes de recherche et des institutions du patrimoine culturel pourront la mobiliser. Cette nouvelle exception est aussi limitée dans son champ d'application puisque les opérations de fouilles de textes et de données doivent être effectuées à des fins de recherche scientifique. Outre les conditions tenant aux bénéficiaires et au but de la fouille de textes et données, le bénéficiaire doit également avoir eu un accès licite au contenu en question. Il n'est également pas possible d'écarter l'exception de l'article 3 par voie contractuelle.

⁴ Dir. (UE) 2019/790 du Parlement européen et du Conseil du 17 avr. 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les dir. 96/9/CE et 2001/29/CE (ci-après DAMUN), JOUE n° L130 du 17 mai 2019.

⁵ S. Le Cam et al., « IA génératives de contenus : pour une obligation de transparence des bases de données ! », *Dalloz Actualité*, 11 mai 2023, en ligne sur : <https://www.dalloz-actualite.fr/node/ia-generatives-de-contenus-pour-une-obligation-de-transparence-des-bases-de-donnees>, consulté le 30 juill. 2024. V. également, V.-L. Benabou, Du test en trois étapes au domaine public payant- Quelques idées pour mieux associer les titulaires de droit à la production des Intelligences Artificielles génératives dans le champ de la création intellectuelle, 11 déc. 2023, Le « chat » et la souris, en ligne sur : <https://vlbenabou.blog/2023/12/11/168/>, consulté le 30 janv. 2024.

⁶ Pour une analyse technique et l'importance des opérations de fouille de textes et de données dans le contexte de l'IA, v. A. Dermawan, "Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment" purposes?", *The Journal of World Intellectual Property*, 2024;27:44–68..

⁷ Art. 2 de la DAMUN.

⁸ Considérant 8 de la DAMUN.

⁹ V. à ce sujet, T. Margoni et al., « A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology », *GRUR International*, 1 août 2022, vol. 71, n° 8, p. 685-701.

¹⁰ J. Han et al., *Data Mining: Concepts and Techniques*, 3rd edition éd., Morgan Kaufmann, 6 juill. 2011.

¹¹ Art. 3 de la DAMUN.

L'article 4 quant à lui concerne tous les bénéficiaires ayant eu un accès licite au contenu objet de la fouille. De plus, il n'y a aucune restriction concernant l'objet des activités de fouille de textes et de données, ce qui signifie que toutes les activités de fouille de textes et de données peuvent bénéficier de cette exception, qu'elles soient exercées à des fins lucratives ou non, pour autant que les autres conditions de l'exception soient respectées. Enfin, une différence majeure avec l'exception de l'article 3 est la possibilité pour les titulaires de droit de réserver leurs droits, grâce au paragraphe 3 de l'article 4. L'article 4, paragraphe 3 dispose que « l'exception ou la limitation prévue au paragraphe 1 s'applique à condition que l'utilisation des œuvres et autres objets protégés visés audit paragraphe n'ait pas été expressément réservée par leurs titulaires de droits de manière appropriée, notamment par des procédés lisibles par machine pour les contenus mis à la disposition du public en ligne. »¹² La directive prévoit ainsi la possibilité pour les titulaires de droits de s'opposer à l'utilisation de leur contenu pour les opérations de fouilles de textes et données, aussi appelé « *opt-out* ». Il s'agit par conséquent d'un système dans lequel l'auteur est invité à se manifester d'une manière ou d'une autre pour s'opposer à l'exploitation de son œuvre¹³.

Les articles 3 et 4 de la DAMUN ont souvent été comparés aux exceptions consacrées dans de nombreux pays, qui semblaient offrir un cadre plus souple. A titre d'exemple, Singapour, Israël ou encore les États-Unis sont fréquemment cités. Toutefois,

l'exception la plus souvent évoquée est l'exception japonaise, considérée par beaucoup comme plus souple et offrant un cadre réglementaire très favorable aux développeurs de systèmes d'IA. Cet article est l'occasion de revenir sur ces affirmations et d'aborder ainsi les conditions de l'exception de fouille de textes et de données japonaise (I) et, notamment, sur sa prétendue flexibilité à l'ère de l'IA générative (II).

I. Les conditions de l'exception de fouille de textes et de données japonaise

Le Japon a souvent été présenté comme un « paradis pour le machine learning », selon le Professeur Tatsuhiro Ueno¹⁴ ou encore Yann LeCun¹⁵. Le droit d'auteur japonais autorise les techniques d'apprentissage automatique depuis 2009 et a été le premier pays au monde à adopter une législation pertinente pour permettre la fouille de textes et de données¹⁶. Cette exception est perçue comme beaucoup plus souple que la directive DAMUN et comme l'une des exceptions les plus permissives.¹⁷ La loi japonaise sur le droit d'auteur vise à établir un équilibre entre la protection des droits et des intérêts des détenteurs de droits d'auteur et l'exploitation harmonieuse des œuvres protégées par le droit d'auteur¹⁸.

Dès 2016, le Japon a identifié l'IA comme une priorité. Pour garantir plus de sécurité juridique et pour offrir une certaine souplesse aux innovateurs en matière d'IA, le gouvernement a publié le 9 mai 2016 son

¹² Art. 4 (3) de la DAMUN.

¹³ J.-M. Bruguière, « Intelligence artificielle et droit d'auteur - Sortir de la science-fiction des « machines/auteurs », entrer dans la réalité du droit des données », *Communication Commerce électronique* 2020, n° 6, étude 11.

¹⁴ T. Ueno, « コラム : 機械学習パラダイス (上野達弘) », *RCLIP - Waseda University*, en ligne sur : <https://www.rclip.jp/publications/2017/09/09/201708column>, consulté le 30 juill. 2024.

¹⁵ V. par exemple ce [tweet](#) de Yann LeCun du 1^{er} juin 2023.

¹⁶ *Japan amends its copyright legislation to meet future demands in AI*, 3 sept. 2018, European Alliance for Research Excellence, en ligne sur : <https://eare.eu/japan-amends-tdm-exception-copyright/>, consulté le 6 mars 2021

¹⁷ T. Ueno, "Conference: Challenges of the Directive 2019/790 on Copyright and Related Rights in the Digital Single Market Law Faculty", Jagiellonian University, Kraków, 25 octobre 2019.

¹⁸ Agency for Cultural Affairs, Government of Japan, General Understanding on AI and Copyright in Japan - Overview, Mai 2024 (ci-après le Document).

programme stratégique pour la propriété intellectuelle, qui propose notamment d'introduire des dispositions « souples » sur les limitations du droit d'auteur afin de promouvoir les nouvelles innovations numériques¹⁹. Lors de la construction de ces exceptions, il est expliqué que plusieurs options ont été étudiées²⁰ comme la possibilité d'introduire des dispositions très flexibles similaires aux dispositions du *fair use* américain²¹. Il a été conclu que l'approche souhaitable serait de combiner plusieurs dispositions avec un équilibre approprié entre clarté et flexibilité, plutôt que d'adopter des dispositions similaires aux dispositions sur le *fair use*²².

Le gouvernement japonais a adopté le projet de loi visant à modifier la loi japonaise sur le droit d'auteur le 18 mai 2018, entré en vigueur le 1^{er} janvier 2019.²³ Cette modification introduit plusieurs nouveaux articles, tels que l'article 30-4, l'article 47-4 ou l'article 47-5.

Le nouvel article 47-4²⁴ reconnaît que la réalisation de copies électroniques accidentelles d'œuvres est importante pour les activités d'apprentissage automatique,

mais n'affecte pas les titulaires de droits d'auteur. Le nouvel article 47-5²⁵, quant à lui, permet aux chercheurs d'utiliser des œuvres protégées par le droit d'auteur pour la vérification des données dans le cadre d'une étude, en reconnaissant qu'une telle utilisation est bénéfique pour les chercheurs et ne porte pas préjudice aux titulaires de droits. Cet article donne accès à des bases de données consultables, qui sont nécessaires pour la vérification des données des résultats de l'apprentissage automatique et des connaissances²⁶.

L'article 30-4 sera l'objet de cette étude puisqu'il s'agit de la disposition la plus intéressante concernant les opérations de fouille de textes et de données.

L'article 30-4 est tout d'abord intitulé « Exploitation sans but de jouir des pensées ou des sentiments exprimés dans une œuvre »²⁷. L'article 30-4 précise qu'il est possible d'exploiter une œuvre pour réaliser des activités de fouilles de textes et de données tant que l'opération entre dans une des quatre activités prévues, à savoir l'extraction, la comparaison, la classification ou d'autres analyses statistiques.²⁸

¹⁹ [Programme stratégique pour la propriété intellectuelle 2016](#), 9 mai 2016.

²⁰ *Regarding the Act Partially Amending the Copyright Act (Act No. 30 of 2018)*, en ligne sur : https://www.bunka.go.jp/english/policy/copyright/amendments_2018/.

²¹ V. à ce sujet, P. Kamina, « Fair use », 2001.

²² *Regarding the Act Partially Amending the Copyright Act (Act No. 30 of 2018)* (préc. note 20).

²³ V., Act No. 30 du 25 Mai 2018. V. également, *Regarding the Act Partially Amending the Copyright Act (Act No. 30 of 2018)* (n20). V. également, Copyright Research and Information Center (CRIC), *Copyright Law of Japan*, en ligne sur : <https://www.cric.or.jp/english/clj/cl2.html>, consulté le 30 juill. 2024.

²⁴ V. Copyright Law of Japan (préc. note 23).

²⁵ V. Copyright Law of Japan (préc. note 23).

²⁶ V. Japan amends its copyright legislation to meet future demands in AI (n 16).

²⁷ Texte original : "Exploitation without the Purpose of Enjoying the Thoughts or Sentiments Expressed in a Work", traduction libre.

²⁸ Article 30-4 de la loi sur le droit d'auteur japonais (Japan Copyright Act), disposant que "It is permissible to exploit a work, in any way and to the extent considered necessary, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work; provided, however, that this does not apply if the action would unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation: (i) if it is done for use in testing to develop or put into practical use technology that is connected with the recording of sounds or visuals of a work or other such exploitation; (ii) if it is done for use in data analysis (meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data; the same applies in Article 47-5, paragraph (1), item (ii)); and (iii) if it is exploited in the course of computer data processing or otherwise exploited in a way that does not involve what is expressed in the work being

De plus, il faut que l'exploitation ne permette pas « de jouir personnellement ou de faire jouir une autre personne des pensées ou des sentiments exprimés dans cette œuvre » et que cela ne porte pas un préjudice injustifié aux intérêts du titulaire du droit d'auteur²⁹. Ainsi, il apparaît que l'exception japonaise peut s'appliquer à des fins commerciales et non commerciale, à toute exploitation quelles que soient les réserves des titulaires de droits, l'exploitation par tous les moyens est autorisée et aucun accès licite n'est requis, à la différence de l'exception européenne par exemple³⁰. C'est donc pour cela que l'exception japonaise est considérée comme étant extrêmement souple en comparaison par exemple de l'article 3 qui est restreint dans sa finalité et ses bénéficiaires, ou de l'article 4 qui nécessite un accès licite à l'œuvre et qui permet un mécanisme d'opposition des titulaires de droits.

Cette exception introduit une notion très intéressante, qui est celle de « jouir de l'œuvre ». Jouir de l'œuvre renvoie ici à « accepter et apprécier des choses hautement émotionnelles ou un intérêt physique » selon l'office du droit d'auteur japonais³¹.

L'office ajoute que « les actes qui ne visent pas à profiter des idées ou des émotions exprimées dans une œuvre ne portent pas

atteinte aux possibilités du titulaire du droit d'auteur de recevoir des compensations de la part de ceux qui ont l'intention de jouir des idées ou des émotions exprimées dans l'œuvre et les intérêts du titulaire du droit d'auteur qui doivent être protégés par la loi sur le droit d'auteur ne seront normalement pas lésés »³². Les actes de fouilles de textes et de données ne sont pas perçus comme permettant de profiter des idées ou des émotions exprimées dans un sens humain.

II. La flexibilité de l'exception japonaise repensée à l'ère de l'IA générative

Si l'exception de fouille de textes et de données japonaise apparaît comme étant très souple, cette affirmation tend à être remise en question et semble plus modérée à l'égard de l'IA générative.

En 2023, un comité de l'Agence japonaise des affaires culturelles a rédigé un document (ci-après le Document) destiné à clarifier l'interaction entre la loi japonaise sur le droit d'auteur et l'IA générative. Pour cela, le comité a recueilli près de 25 000 commentaires de juristes, créateurs, ingénieurs, ainsi que d'entreprises de tous secteurs et disciplines. Le 15 mars 2024, la version finale du document a été publiée. Tout d'abord, le rapport souligne que normalement, les dispositions de la loi sur le

perceived by the human senses (for works of computer programming, such exploitation excludes" Traduction libre: « Il est permis d'exploiter une œuvre, de quelque manière et dans la mesure jugée nécessaire, dans l'un des cas suivants, ou dans tout autre cas où il n'est pas dans l'intention d'une personne de profiter personnellement ou de faire profiter une autre personne des pensées ou des sentiments exprimés dans cette œuvre ; à condition, toutefois, que cela ne s'applique pas si l'action porterait un préjudice déraisonnable aux intérêts du titulaire du droit d'auteur compte tenu de la nature ou de l'objectif de l'œuvre ou des circonstances de son exploitation : (i) si cela est fait pour une utilisation dans des tests visant à développer ou à mettre en pratique une technologie liée à l'enregistrement de sons ou d'images d'une œuvre ou à toute autre exploitation similaire ; (ii) si cela est fait pour une utilisation dans l'analyse de données (c'est-à-dire l'extraction, la comparaison, la classification ou toute autre analyse statistique des éléments

constitutifs de la langue, des sons, des images ou d'autres données élémentaires provenant d'un grand nombre d'œuvres ou d'un grand volume d'autres données similaires ; il en va de même à l'article 47-5, paragraphe (1), point (ii)) ; et (iii) si cela est exploité dans le cadre du traitement de données informatiques ou autrement exploité d'une manière qui n'implique pas que ce qui est exprimé dans l'œuvre soit perçu par les sens humains (pour les œuvres de programmation informatique, une telle exploitation exclut ».

²⁹ *Ibid.*

³⁰ A. Dermawan, „Text and data mining exceptions in the development of generative AI models”, *préc.*

³¹ Japan Copyright Office (JCO), Outline of the Amendments to the Copyright Act in 2018, Patents & Licensing,

Avril 2019.

³² *Ibid.*, p.12

droit d'auteur doivent être analysées à la lumière de la jurisprudence et donc, au cas par cas. Toutefois, il existe actuellement que très peu de décisions judiciaires permettant de dégager une interprétation de la loi et traitant directement de l'interaction entre l'IA et la loi sur le droit d'auteur japonais.

Il est à noter que ce document n'est pas juridiquement contraignant. Ce document vise à clarifier certaines situations tenant à l'application de l'exception. L'étude du Document se concentrera ici sur les précisions apportées concernant les phases amont, à savoir l'entraînement et les opérations de fouilles de textes et de données³³.

Tout d'abord, un des apports important du Document concerne la notion de jouissance de l'œuvre. Le Document souligne que dans certaines situations, il peut y avoir une coexistence du but de jouissance et de non-jouissance de l'œuvre, rendant l'article 30-4 inapplicable. Pour déterminer s'il y a une jouissance de l'œuvre, il faut analyser l'œuvre utilisée au sens de l'article. Il ne faut pas analyser d'autres œuvres protégées par le droit d'auteur ou le style d'un artiste. Le Document précise également qu'une autorisation peut être nécessaire de la part du titulaire des droits lorsqu'une œuvre est utilisée à des fins « non commerciales » ou « de recherche » s'il existe également un « but

de jouissance » de l'œuvre. En outre, l'article 30-4 n'aura pas vocation à s'appliquer dans les situations qui « porteraient un préjudice déraisonnable aux intérêts du titulaire du droit d'auteur »³⁴.

Le Document détaille plusieurs exemples où l'on peut retrouver cette coexistence. Par exemple lors de l'exploitation d'œuvres pour l'apprentissage des systèmes d'IA et notamment lors de la collecte des données d'apprentissage. En effet, cette coexistence peut exister lorsque certaines œuvres sont collectées pour entraîner un système d'IA qui générera du contenu similaire aux œuvres présentes dans les données d'entraînement³⁵. Il s'agit ici d'une forme d'entraînement supplémentaire sur le modèle de fondation, également connu sous le nom d'ajustement fin (*fine-tuning*³⁶) et de « surajustement » intentionnel (*overfitting*³⁷)³⁸.

Le Document analyse une des problématique phare des systèmes d'IA générative, à savoir la reproduction d'un contenu imitant le style d'un artiste. Il s'agit ici d'un sujet omniprésent dans les discussions autour de l'IA et du droit d'auteur puisque de nombreux titulaires de droits regrettent la production de contenus semblables à leur style, sans toutefois reproduire de façon *verbatim* des éléments protégés. Le Document rappelle, et ce comme c'est le cas traditionnellement, que le droit d'auteur ne

³³ Le Document évoque également d'autres sujets tels que l'atteinte au droit d'auteur dans les contenus générés ainsi que l'utilisation de tels contenus mais aussi la protection par le droit d'auteur des contenus générés.

³⁴ General Understanding on AI and Copyright in Japan – Overview (*préc. note 18*), page 6. L'avis prend l'exemple de la reproduction d'une base de données protégée par le droit d'auteur à des fins d'analyse de données, comme l'entraînement de systèmes d'IA pour lequel des licences d'analyse de données sont disponibles sur le marché. La notion de préjudice déraisonnable aux intérêts du titulaire du droit d'auteur est précisée par le document, *v. infra*.

³⁵ General Understanding on AI and Copyright in Japan – Overview (*préc. note 18*), page 8.

³⁶ L'ajustement fin, ou *fine-tuning*, est une « technique consistant à spécialiser un modèle d'IA pré-entraîné à l'accomplissement d'une tâche spécifique. Cela

consiste généralement à entraîner le modèle dans son ensemble, ou seulement certaines couches d'un réseau de neurones, pour un faible nombre d'itérations sur un ensemble de données spécifiques correspondant à la tâche visée », *v. à ce sujet Ajustement (fine-tuning)*, en ligne sur : <https://www.cnil.fr/fr/definition/ajustement-fine-tuning>, consulté le 31 juill. 2024.

³⁷ Le surapprentissage, ou *overfitting*, désigne la situation lorsque le modèle ne se généralise pas bien à de nouvelles données non encore connues du modèle. Le modèle fonctionne parfaitement sur l'ensemble de données d'apprentissage mais ne parvient pas à s'ajuster à des données inconnues. *V. à ce sujet, X. Ying, « An Overview of Overfitting and its Solutions », Journal of Physics: Conference Series, 1 févr. 2019, vol. 1168, p. 022022.*

³⁸ General Understanding on AI and Copyright in Japan – Overview (*préc. note 18*), page 8.

protège pas les idées en tant que telles. Il faut une matérialisation de l'idée. Cela signifie donc que le style d'un artiste n'est pas protégé en tant que tel par le droit d'auteur. Ainsi, un contenu généré par un système d'IA qui reproduit le style d'un créateur d'une œuvre préexistante protégée par le droit d'auteur n'enfreint pas le droit d'auteur si le style englobe simplement une idée³⁹.

Toutefois, après avoir rappelé ce principe, le Document présente des scénarios où malgré tout, l'article 30-4 n'aura pas vocation à s'appliquer en cas de reprise du style d'un artiste. Par exemple, si un groupe d'œuvres est intentionnellement reproduit afin d'effectuer un entraînement supplémentaire dans le but de générer des contenus qui contiennent tout ou partie de ces expressions créatives communes, l'article 30-4 de la loi ne s'applique pas à une telle reproduction car le « but de jouissance » coexiste⁴⁰.

Le Document revient également sur la clause restrictive de l'article 30-4 qui permet l'exploitation de contenu protégé par le droit d'auteur sauf en cas de préjudice déraisonnable aux intérêts du titulaire de droits. Pour que cette clause restrictive s'applique, il faudra analyser si une concurrence à l'œuvre protégée sur le marché est observée et si les circuits de vente futurs potentiels de l'œuvre protégée seront impactés. Cette évaluation doit être effectuée en tenant compte de divers facteurs, tels que les « progrès technologiques » et les « changements dans la manière dont l'œuvre protégée est utilisée »⁴¹. Le rapport explique ensuite la relation entre la collecte de données d'entraînement à l'IA et cette clause restrictive.

Un des arguments avancés par la doctrine pour justifier de l'apparente souplesse de l'exception japonaise était qu'il n'était vraisemblablement pas nécessaire d'avoir un accès légal à l'œuvre utilisée pour pouvoir en bénéficier⁴². Toutefois, cette affirmation tend à être nuancée par le Document. Tout d'abord, il est indiqué que « la collecte de données d'entraînement à l'IA sur des sites web dont on sait qu'ils diffusent des contenus piratés ou violant le droit d'auteur doit être strictement évitée »⁴³.

Le Document explique ainsi que si une entreprise, c'est-à-dire un développeur de systèmes d'IA ou un prestataire de services de systèmes d'IA, collecte sciemment des données d'entraînement qui comprennent des copies contrefaisantes, cette entreprise peut être tenue pour responsable d'une atteinte au droit d'auteur par l'IA générative dans certains cas. Selon la jurisprudence, il est possible d'être responsable d'une atteinte au droit d'auteur tout en n'étant pas le responsable direct de l'atteinte. Cela signifie par exemple qu'un développeur d'IA ou un prestataire de services d'IA collectant des données d'entraînement à partir d'un site web dont il a connaissance qu'il contient des contenus piratés ou contrefaits peut être responsable d'une atteinte au droit d'auteur. Bien que l'entreprise ne soit pas directement responsable de l'infraction puisqu'elle utilise des copies contrefaisantes sans avoir effectué l'acte contrefaisant, à savoir la reproduction illicite, elle peut toutefois être tenue responsable de l'atteinte.

Finalement, une dernière remarque intéressante concerne la notion de dépendance entre les données d'entraînement et la génération d'un contenu

³⁹ J.-M. Deltorn, « In the style of... - deep learning, style transfer and the limits of copyright protection. A European perspective », *Annali italiani del diritto d'autore, della cultura e dello spettacolo*, 2019, vol. XXVII, n° XXVII-2020, p. 337.V. également les mesures prises par certaines grandes plateformes mettant à disposition des outils d'IA qui interdisent de générer du contenu lorsque le prompt indique, par exemple, « dans le style de » suivi du nom d'un artiste.

⁴⁰ General Understanding on AI and Copyright in Japan - Overview (*préc. note 18*), page 9

⁴¹ General Understanding on AI and Copyright in Japan - Overview (*préc. note 18*), page 10.

⁴² A. Dermawan, « Text and data mining exceptions in the development of generative AI models », *préc.*

⁴³ General Understanding on AI and Copyright in Japan - Overview (*préc. note 18*), page 11.

par un système d'IA. Bien que cela concerne la génération d'un objet et non la phase de fouille de textes et de données *stricto sensu*, il existe malgré tout un lien puisque cela implique la phase en amont, lors de laquelle la fouille de textes et de données peut avoir vocation à jouer. En effet, il est souvent argué par de nombreux artistes qu'au regard de certains contenus générés, il est plus probable que les données utilisées aient été les données protégées par le droit d'auteur d'artistes⁴⁴.

Le Document explique que, pour constater une violation du droit d'auteur par un contenu généré, il faut que ce contenu ait une « similitude » et une « dépendance » avec une œuvre protégée par le droit d'auteur.

Le Document évoque le processus d'entraînement comme pouvant expliquer la dépendance entre le contenu utilisé pour l'entraînement et le contenu généré. La notion de « dépendance » dans le cas d'un contenu généré par un système d'IA est expliquée. Le Document précise que dans les situations où il n'est pas possible de savoir avec certitude si une œuvre protégée a été utilisée lors de la phase d'apprentissage, il sera présumé une « dépendance » si le titulaire du droit d'auteur peut prouver que « l'utilisateur de l'IA a eu accès à l'œuvre protégée existante » ou que « le matériel généré par l'IA présente un degré élevé de similitude avec l'œuvre ». Cela signifie qu'il est possible pour le titulaire du droit d'auteur d'établir une dépendance même s'il n'est pas établi avec certitude qu'une œuvre a été utilisée lors de l'entraînement du système d'IA. Il s'agit ici d'une position très intéressante et méritant d'être soulignée puisque les titulaires de droits se retrouvaient souvent confrontés au problème suivant. Ils

constataient que certains systèmes d'IA générative pouvaient produire des contenus similaires à leurs œuvres protégées mais sans pour autant reproduire de façon *verbatim*⁴⁵ des éléments d'un contenu protégé. De ce fait, difficile pour eux de pourvoir arguer d'une violation du droit de reproduction dans le contenu généré par exemple. De plus, en l'absence de transparence sur les données utilisées pour entraîner le système d'IA, il est presque impossible de prouver l'utilisation de certaines œuvres protégées par le droit d'auteur au moment de l'entraînement. La précision du Document apporterait une solution à ces situations.

Enfin, lorsque le titulaire d'un droit d'auteur peut démontrer qu'une œuvre protégée est utilisée pour la phase d'entraînement, il est généralement admis qu'il existe une dépendance vis-à-vis de cette œuvre préexistante, même si l'utilisateur de l'IA n'en était pas conscient, dès lors que l'œuvre a été employée durant la phase de développement de l'IA⁴⁶. Toutefois, cette précision est tempérée puisque si le contenu généré ne reproduit pas « l'expression protégée » de l'œuvre utilisée lors de l'entraînement, alors il peut être considéré qu'il n'y a pas de dépendance.

Conclusion

Les opérations de fouille de textes et de données ont été l'objet de vives discussions ces dernières années à la suite de l'adoption des articles 3 et 4 de la DAMUN mais également de leur potentiel rôle pour le développement des systèmes d'IA.

Les articles 3 et 4 de la DAMUN n'ont pas toujours fait l'unanimité auprès de la doctrine. Ils ont souvent été perçus comme

⁴⁴ V. à ce sujet, S. Le Cam et al., *Argumentaire de la Ligue : pour une régulation des IA*, 4 avr. 2023, Ligue des auteurs professionnels, en ligne sur : <https://ligue.auteurs.pro/2023/04/04/argumentaire-de-la-ligue-pour-une-regulation-des-ia/>, consulté le 31 juill. 2024.

⁴⁵ V. à ce sujet l'affaire en cours aux États-Unis, *The New York Times Company c. Microsoft Corporation and OpenAI* (1:23-cv-11195), 27 décembre 2023. Cette

affaire a été particulièrement commentée puisque le demandeur fournit des preuves de reproductions *verbatim* de contenus, facilitant la démonstration d'une contrefaçon par la reproduction à l'identique dans le contenu généré, ce qui est particulièrement intéressant en l'absence de transparence sur les données d'entraînement.

⁴⁶ General Understanding on AI and Copyright in Japan – Overview (*préc. note 18*), page 13.

des exceptions mal adaptées aux contextes du développement de la technologie et plus particulièrement des systèmes d'IA. Ces articles ont régulièrement été comparés à des exceptions étrangères et plus spécifiquement à l'exception japonaise.

L'exception japonaise a été présentée comme une exception très large et très favorable aux opérations de fouille de textes et de données. Bien que de prime abord ses conditions sont assez permissives, par exemple grâce au concept de «jouissance de l'œuvre», il apparaît que cette acception selon laquelle l'exception japonaise est extrêmement favorable aux développeurs de systèmes d'IA tend à être modérée, notamment dans le contexte de l'IA générative.

En effet, le récent Document mis à disposition par le comité de l'Agence japonaise des affaires culturelles en 2024 précise plusieurs scénarios où l'exception ne pourrait finalement pas s'appliquer ou de façon limitée. Bien que ce Document ne soit pas juridiquement contraignant et qu'il souligne l'importance de la jurisprudence pour déterminer l'étendue de l'exception japonaise, il se présente comme une première piste pour analyser plusieurs situations dans lesquelles co-existent par exemple une finalité de jouissance de l'œuvre et de non-jouissance de l'œuvre. D'autres thématiques comme l'imitation du style d'un créateur ou de la dépendance entre le contenu utilisé pour l'entraînement et le contenu généré sont également abordées, fournissant des éléments de réflexion très intéressants pouvant être utiles pour une réflexion au niveau européen.

E. M.