Vol. 32, n° 1

Le droit d'auteur canadien à l'ère de l'intelligence artificielle : le cas du traitement du langage naturel

Hélène Beauchemin*

© Hélène Beauchemin, 2020.

* Conseillère juridique (avocate) chez Stradigi AI.

[Note : cet article a été soumis à une évaluation à double anonymat.]

RÉSUMÉ

L'intelligence artificielle bouleverse tous les secteurs et comme toute innovation, elle amène son lot d'incertitudes. Le droit d'auteur dans sa forme actuelle n'est pas adapté aux processus technologiques utilisés par l'intelligence artificielle et constitue un obstacle à la pleine réalisation de son potentiel. Le texte suivant vise à dresser un portrait global des difficultés d'interprétation de certaines dispositions de la *Loi sur le droit d'auteur* appliqué au contexte spécifique du traitement et de la compréhension du langage naturel.

INTRODUCTION

L'expression « intelligence artificielle » (ou « IA ») est de nos jours sur toutes les lèvres. D'une part, les consommateurs désirent bénéficier de ses nombreux avantages ; les entreprises, quant à elles, s'interrogent sur les manières de l'intégrer efficacement dans leurs activités commerciales. D'autre part, cette transformation vers l'industrie 4.0 provoque de nombreuses inquiétudes sur son utilisation et sur les conséquences de celle-ci, certaines semblant à première vue inévitables. On ne commence cependant qu'à entrevoir les répercussions inattendues de l'application de cette technologie. L'une des répercussions moins discutées dans la sphère publique touche à l'étendue de la protection du droit d'auteur et à l'interprétation de ses exceptions.

Cet enjeu fait appel à une notion centrale du droit d'auteur : le droit exclusif de reproduire la totalité ou une partie d'une œuvre[1]. Le hic : le processus d'entraînement des algorithmes, notamment les techniques de traitement et de compréhension du langage naturel et de reconnaissance d'images, nécessite la production d'une copie temporaire afin d'extraire les éléments pertinents à l'analyse et à la production d'un résultat final, soit le modèle entraîné.

Cette situation rend difficile l'application traditionnelle des principes fondamentaux du droit d'auteur, particulièrement en raison du fait que l'intelligence artificielle n'est pas développée en vase clos et dépasse souvent les murs des institutions de recherche[2], rendant donc les exceptions exhaustives prévues par la *Loi sur le*

1. *Loi sur le droit d'auteur*, L.R.C. (1985), c. C-42, (ci-après, la « Loi »), art. 3(1).
2. Un rapport récent publié par l'Organisation mondiale de la propriété intellectuelle révèle que contrairement à d'autres industries, l'innovation en intelligence artificielle est menée majoritairement par l'industrie privée (26 des 30 plus importants demandeurs de brevets étaient des entreprises privées, les quatre derniers étant des instituts d'enseignement), tandis que les universités demeurent actives en recherche dans certains domaines spécifiques : voir WIPO, « WIPO Technology Trends 2019 – Artificial Intelligence », 2019, accessible sur <https://www.wipo.int/publications/en/details.jsp?id=4386, p. 15-16>.

droit d'auteur (la « Loi »)[3], dans sa version actuelle, largement insuffisantes. Le statut juridique incertain de ces modèles, entraînés à partir de bases de données dont les droits sont flous[4] et contenant des modalités de licences manquant également de clarté[5], ainsi que la nature restrictive du cadre de la Loi, causent des barrières à l'innovation[6]. Ces obstacles se manifestent dans plusieurs sphères de l'intelligence artificielle, notamment les voitures autonomes[7] : les fabricants doivent naviguer un paysage juridique marqué, d'un côté, par l'absence de normes dans certaines régions et, de l'autre, par la présence de normes parfois incompatibles, résultant ainsi en une incertitude juridique, et de manière concrète, en un ralentissement considérable de la vitesse d'innovation et d'introduction de nouvelles technologies[8]. Le problème des restrictions propres au droit d'auteur et leurs effets sur l'innovation ne sont d'ailleurs pas limités à l'intel-

3. *Loi sur le droit d'auteur*, *supra*, note 1, art. 29 et s.
4. Kashmir HILL et Aaron KROLIK, « How Photos of Your Kids Are Powering Surveillance Technology », *New York Times*, 11 octobre 2019, accessible sur <https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html> (sur la question de la provenance d'images d'une base de données provenant d'individus n'ayant jamais donné leur consentement à la collecte, ni à l'utilisation de leurs photos de famille à des fins de reconnaissance faciale).
5. À l'exception des bases de données provenant de sociétés privées ou gouvernementales, une grande quantité de base de données utilisées dans l'industrie de l'IA proviennent du milieu académique et ne font mention de modalités minimales, ou dans certains ne font mention d'aucune licence. Quelques exemples : (1) une base de données bâtie pour être utilisée en traitement du langage naturel répertoriant 200 000 questions de l'émission américaine *Jeopardy*, obtenues à l'aide d'une technique d'extraction de contenu (« web scraping »), qui ne fait état d'aucune licence, accessible sur <https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/> (le statut juridique de cette technique étant lui-même contesté à certains égards) ; (2) The Blog Authorship Corpus, accessible sur <http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm> (la seule mention de licence est que le jeu de données doit être utilisé à des fins de recherche non commerciale).
6. Mark FENWICK, Wulf A. KAAL et Erik P.M. VERMEULEN, « Regulation Tomorrow: What Happens When Technology Is Faster Than the Law? », (2017) 6-3 *American University Business Law Journal Review* 561, p. 573 ; Adam THIERER, *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom*, Mercatus Center at George Mason University, 2016, et Deirdre N. MCCLOSKEY, *The Bourgeois Virtues: Ethics for an Age of Commerce*, The University of Chicago Press, 2006 (la notion d'innovation « sans permission » est au cœur de l'argument selon lequel la richesse économique est créée grâce à un système de valeurs permettant l'innovation *de facto*, favorisant ainsi l'expérimentation et tolérant les risques engendrés par celle-ci).
7. Jessica S. BRODSKY, « Autonomous Vehicle Regulation: How an Uncertain Legal Landscape May Hit the Brakes on Self-Driving Cars », (2016) 31 *Berkeley Technology Law Journal* 851, p. 851-878.
8. Jo Ann S. BAREFOOT, « Disrupting Fintech Law », (2015) 18-2 *Fintech Law Report*, p. 10 : un autre exemple connexe, mais qui n'est pas restreint à l'intelligence artificielle, se situe dans le domaine de la technologie financière, un domaine hautement réglementé : l'entrée en marché de plusieurs produits peut avoir lieu avant même que la législation pertinente n'ait pu être modifiée en vertu du processus législatif.

ligence artificielle, mais à d'autres pratiques émergentes, comme la création de contenu généré par les utilisateurs[9]. En effet, dans ce cas, certains auteurs mentionnent que les restrictions en droit d'auteur restreignent le développement du plein potentiel de cette pratique et, par le fait même, les droits des utilisateurs, en raison des risques de litige importants, particulièrement par des collectifs représentant des groupes d'auteurs[10]. On peut facilement envisager un effet similaire potentiel avec l'usage de bases de données contenant des quantités massives de données protégées par la Loi.

Certains commentateurs argumentent de façon plus spécifique que les restrictions inhérentes au droit d'auteur sont au cœur de la friction entre le développement de l'IA et son utilisation à grande échelle, entraînant ainsi des problèmes cruciaux : on peut penser notamment aux biais implicites des concepteurs humains[11], reproduits dans les bases de données faciles d'accès, ou encore, à des bases de données dont la composition demeure secrète. Lorsque ces bases de données, d'une qualité variable, sont utilisées pour entraîner des algorithmes servant à prendre des décisions ayant un effet sur des individus, le choix de ces bases de données peut donc avoir des conséquences importantes sur la vie de ceux-ci[12]. Cependant, faute d'accès à des ressources alternatives, plusieurs joueurs doivent se contenter d'un accès limité aux bases de données pertinentes et de qualité, lesquelles sont difficile d'accès[13] et, de surcroît, font potentiellement

9. Samuel TROSOW, « Copyright as Barrier to Creativity: The Case of User-Generated Content », *Intellectual Property for the 21st Century: Interdisciplinary Approaches*, Irwin Law, 2014, p. 521 et s.
10. *Ibid.*, p. 529.
11. Amanda LEVENDOWSKI, « How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem », (2018) 93 *Washington Law Review* 579, p. 589 (plus précisément, l'auteure argumente que l'application de l'exception du *fair use* prévu par le droit d'auteur américain peut amener à la réduction de ce biais) [*LEVENDOWSKI*] ; le même argument est soutenu par le professeur Ryan Calo de l'Université de Washington, voir Ryan CALO, « Artificial Intelligence Policy: A Primer and Roadmap », (2017) 51 *University of California Davis Law Review* 399, p. 424.
12. Julia ANGWIN, Jeff LARSON, Surya MATTU et Lauren KIRCHNER, « Machine Bias », 23 mai 2016, accessible sur <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (c'est l'argument d'un rapport de l'organisation ProPublica, qui a révélé dans une enquête en 2016 qu'un des problèmes des analyses de risques de récidive d'infractions criminelles réalisées par des entreprises privées est que celles-ci protègent jalousement non seulement les méthodes pour arriver aux résultats, mais également les données utilisées, rendant ainsi impossible la contestation du bien-fondé des décisions rendues par les systèmes de prédictions).
13. Darrell M. WEST et John R. ALLEN, « How artificial intelligence is transforming the world », Brookings Institute, 24 avril 2018, accessible sur <https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/>

l'objet de la protection de la Loi lorsqu'elles répondent aux exigences de celle-ci[14].

Il est donc aisé de concevoir qu'en pratique, ces données demeurent hors de l'accès de la majorité des joueurs de l'IA, à l'exception des entreprises dominantes[15]. D'autre part, ces acteurs principaux ont peu d'incitatifs à rendre publiques les données utilisées, car au-delà de l'avantage concurrentiel, révéler les sources de celles-ci pourrait engendrer des enjeux de responsabilité en vertu du droit d'auteur[16].

Malgré ce qui précède, il est possible d'entrevoir une solution à tout le moins partielle aux défis décrits aux présentes. En effet, le Comité permanent de l'industrie, des sciences et de la technologie (« Comité »), à la suite de consultations avec des témoins provenant de divers secteurs d'activités, a reconnu l'importance de moderniser la Loi et a de ce fait recommandé dans son plus récent rapport d'examen que l'exception de l'utilisation équitable soit modifiée afin de prendre en considération la réalité pratique des techniques utilisées en intelligence artificielle[17].

Ce texte vise à fournir au lecteur un bref portrait des méthodes d'intelligence artificielle visées, les dispositions législatives pertinentes et la jurisprudence applicable, ainsi qu'à mettre en lumière les incompatibilités résultant de l'écart entre la réalité technologique et la Loi. Finalement, des suggestions seront proposées afin de pro-

---

(un rapport identifiant l'accès aux données comme étant un problème limitant l'innovation et le design des systèmes et dont la première recommandation consiste à améliorer l'accès aux données).

14. *Loi sur le droit d'auteur*, *supra*, note 1, art. 5(1) (l'œuvre doit être matérialisée et originale) ; *Théberge* c. *Galerie d'Art du Petit Champlain inc.*, 2002 CSC 34, par. 25 [*Théberge*] (l'œuvre doit être également fixée et posséder un caractère permanent) ; *CCH Canadienne Ltée* c. *Barreau du Haut-Canada*, 2004 CSC 13, par. 25 [*CCH*] (l'œuvre doit faire l'objet de « l'exercice du talent et du jugement »).
15. *LEVENDOWSKI*, *supra*, note 11, p. 597 et 599 (la dominance d'une poignée d'entreprises multinationales en IA, telles que Google, Apple, Baidu, et Microsoft, rend difficile l'entrée sur le marché de nouveaux concurrents, lesquels doivent composer avec l'écart de ressources nécessaires pour convertir de nouveaux clients ou encore mettre en œuvre des techniques de gestion des biais ; cela est d'autant plus vrai que bien que plusieurs algorithmes soient disponibles en code source libre, il est rare que les bases de données ayant servi à la création de ceux-ci le soient).
16. Voir généralement Frank PASQUALE, *The Black Box Society: Algorithms That Control Money and Information*, Harvard University Press, 2015.
17. COMITÉ PERMANENT DE L'INDUSTRIE, DES SCIENCES ET DE LA TECHNOLOGIE (CHAMBRE DES COMMUNES DU CANADA), « Examen prévu par la loi de la *Loi sur le droit d'auteur* », déposé le 3 juin 2019, Recommandations 18, 19 et 23.

poser des avenues législatives possibles pouvant résoudre certaines difficultés actuelles auxquelles font face les acteurs de l'IA.

1. L'INTELLIGENCE ARTIFICIELLE : TERMINOLOGIE ET DÉFINITIONS

Cette première section vise à définir de façon sommaire les principaux termes et les concepts pertinents pour la compréhension des défis juridiques posés par le traitement et la compréhension du langage naturel.

1.1 La portée de la notion d'intelligence artificielle

Il est nécessaire de clarifier d'emblée la définition de l'expression « intelligence artificielle ». Dans sa forme la plus simple, on peut décrire cette notion de la façon suivante :

> Domaine d'étude ayant pour objet la reproduction artificielle des facultés cognitives de l'intelligence humaine dans le but de créer des systèmes ou des machines capables d'exécuter des fonctions relevant normalement de celle-ci.[18]

L'intelligence artificielle inclut donc une panoplie de techniques visant à s'approcher du fonctionnement du cerveau humain, qui sont de l'ordre de la compréhension, de la perception, ou de la décision, et qui tentent d'imiter, par exemple, les réseaux de neurones.

Historiquement, l'intelligence artificielle classique visait à modéliser la résolution d'un problème en utilisant des raisonnements logiques[19]. La popularité fulgurante des techniques d'intelligence artificielle dans la dernière décennie s'explique, d'une part, par la puissance de calcul et l'augmentation de la capacité de stockage des ordinateurs, mais, d'autre part, par leur approche moderne, basée sur l'induction plutôt que la déduction[20]. Cette approche permet l'analyse de vastes quantités de données disponibles sans avoir besoin de spécifier des règles préétablies, ni d'utiliser des données non structurées. De plus, elle se fonde sur les connaissances qui peuvent être dérivées

18. OFFICE QUÉBÉCOIS DE LA LANGUE FRANÇAISE, « Intelligence artificielle », 2017, accessible sur <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8385376>.
19. C'est notamment le cas des systèmes expert.
20. Brian Jack COPELAND, « Artificial Intelligence », *Encyclopédie Britannica*, 20 juillet 1998 (dernières révisions datant de mai 2019), accessible sur <https://www.britannica.com/technology/artificial-intelligence/Methods-and-goals-in-AI#ref219086>.

de grandes quantités d'information, plutôt que sur les compétences d'un expert, permettant ainsi des gains d'efficacité importants en augmentant la quantité d'éléments d'analyse, maximisant ainsi l'apport humain en le combinant à la puissance de calcul.

Une spécialisation de l'intelligence artificielle, l'apprentissage automatique (*machine learning*), permet à un programme d'acquérir des connaissances et des aptitudes nouvelles et d'améliorer son efficacité, en se fondant sur les résultats obtenus lors de traitements précédents[21].

Il est à noter que malgré les avancées rapides de la recherche en intelligence artificielle, l'état actuel de la technologie est encore bien loin des robots quasi humains dont on fait le portrait dans la science-fiction, qui posséderaient une intelligence artificielle « générale » ou « forte »[22]. En effet, même la réplication du système nerveux d'un simple ver par un ordinateur (dont la connaissance de l'anatomie est bien connue) n'a pas encore été rendue possible[23]. Force est de constater que la création d'une machine humaine fait donc partie d'un horizon lointain.

Par opposition à cette intelligence artificielle générale, l'intelligence artificielle « étroite » ou « faible »[24], vise à accomplir une ou plusieurs tâches précises. Dans le cas du traitement et de la compréhension du langage naturel, ces tâches permettent, particulièrement lorsqu'elles sont combinées entre elles, d'obtenir des résultats riches en informations, comme nous le verrons ci-dessous. Les techniques de pointe en traitement et en compréhension du langage naturel reposent en grande partie sur l'apprentissage automatique.

1.2 Le traitement et la compréhension du langage naturel

Le traitement du langage naturel consiste en une branche de la linguistique informatique s'intéressant aux interactions entre les

21. OFFICE QUÉBÉCOIS DE LA LANGUE FRANÇAISE, « Apprentissage automatique », 2017, accessible sur <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8395061>.
22. B.J. COPELAND, *supra*, note 20, sous le paragraphe « Methods and Goals in AI ». On y réfère parfois également à une intelligence artificielle « forte ».
23. Voir le projet *OpenWorm* : accessible sur <http://openworm.org/>.
24. OFFICE QUÉBÉCOIS DE LA LANGUE FRANÇAISE, « Intelligence artificielle faible », 2017, accessible sur <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=26543874>.

ordinateurs et les langages dits « naturels »[25], soit les langages utilisés par les humains dans leurs communications entre eux.

Les objectifs généraux théoriques de la linguistique informatique consistent en l'élaboration de cadres grammaticaux et sémantiques permettant l'utilisation d'informations par un ordinateur, la découverte de techniques de traitement et d'apprentissage exploitant les caractéristiques statistiques du langage, ainsi que le développement de modèles informatiques pouvant imiter de façon plausible l'apprentissage ayant lieu dans le cerveau humain[26]. Les applications de la linguistique informatique peuvent être aussi variées que la réponse à des questions (lesquelles peuvent être de simples questions factuelles ou encore des questions comportant un degré de complexité plus élevé, comme des réponses requérant des inférences, des descriptions ou des explications), la production de résumés de textes, ou encore l'analyse de textes ou de langages oraux sur un sujet particulier afin d'analyser les sentiments ou d'autres attributs psychologiques de ceux-ci. Les premières approches de linguistique informatique utilisant l'intelligence artificielle remontent aux années 1960[27]. Cependant, l'approche moderne statistique dominante, centrée sur les corpus (décrite ci-dessous), date plutôt des années 1990[28].

Il est à noter qu'il existe une distinction entre les notions de « traitement » et de « compréhension » du langage naturel. Tout d'abord, la première englobe la deuxième et consiste en la *transformation* d'un texte en données structurées, tandis que la deuxième vise à fournir une compréhension du *sens* des données, en fonction de la grammaire, du contexte, permettant ainsi d'en dériver une intention (les verbes ou les activités incluses dans la phrase) et un contenu (les

25. Cela étant par opposition aux langages de programmation, lesquels sont qualifiés de langage « formels ». OFFICE QUÉBÉCOIS DE LA LANGUE FRANÇAISE, « Langage naturel », 2018, accessible sur <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8372907> et « Langage formel », 1990, accessible sur <http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8408168>. Voir également « Natural Language », *Cambridge Dictionary*, accessible sur <https://dictionary.cambridge.org/dictionary/english/natural-language>.
26. Lenhart SCHUBERT, « Computational Linguistics », (printemps 2019) *The Stanford Encyclopedia of Philosophy*, Edward N. ZALTA (dir.), accessible sur <https://plato.stanford.edu/archives/spr2019/entries/computational-linguistics/>.
27. Voir notamment : Marvin MINSKY, *Semantic Information Processing*, Cambridge, MIT Press, 1968 ; Roger C. SCHANK et Kenneth M. COLBY, *Computer Models of Thought and Language*, San Francisco : W.H. FREEMAN and Co., 1973.
28. Voir à cet effet : John F. ALLEN, *Natural Language Understanding*, Redwood City, Benjamin/Cummings, 1995 ; Daniek JURAFSKY et James H. MARTIN, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2e éd., Prentice-Hall, 2009.

noms ou le contenu d'une action spécifique)[29]. Aux fins du présent texte, nous nous limiterons à un exemple spécifique d'application du traitement du langage naturel, soit la fouille de textes (*text mining*)[30]. Plus précisément, les enjeux juridiques seront illustrés grâce à un flux des travaux typique utilisé.

On peut définir la fouille de textes de la façon suivante :

> [...] la découverte (à l'aide d'outils informatiques) de nouvelles informations en extrayant différentes données provenant de plusieurs documents textuels. Un élément fondamental de ce processus réside dans les relations identifiées entre les informations extraites afin d'identifier de nouveaux faits ou de nouvelles hypothèses à explorer.[31]

La fouille de textes vise à réaliser quatre tâches principales, soit (1) la recherche d'information ; (2) la classification ; (3) l'annotation ; et (4) l'extraction d'informations. La notion de tâche a un but précis et peut être spécifiée par ses données entrantes, ses données sortantes, ainsi que les ressources et les programmes qui sont sollicités dans les circonstances.

De façon sommaire, afin de traiter et de comprendre un langage naturel, un algorithme doit être entraîné. Il doit à cet effet ingérer un certain volume d'information, sélectionné en fonction du but indiqué (par exemple la pertinence, la langue, le registre de celle-ci, etc.), appelé corpus, lequel devra suivre un processus de prétraitement pour ensuite être analysé. Notons qu'une distinction doit être effectuée entre la notion d'« algorithme », soit le programme informatique qui effectue les étapes prédéterminées prévues et décrites dans son code source, et celle de « modèle », lequel résulte de l'entraînement de cet algorithme à l'aide d'un jeu de données particulier.

Afin de construire un modèle, cinq étapes sont généralement nécessaires[32] : (1) la compilation du corpus ; (2) le prétraitement du

---

29. Chethan KUMAR GN, « NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part-1) », Towards Data Science, 25 septembre 2018, accessible sur <https://towardsdatascience.com/nlp-vs-nlu-vs-nlg-know-what-you-are-trying-to-achieve-nlp-engine-part-1-1487a2c8b696>.
30. La fouille de textes est une héritière de la fouille de données (*data mining*) : en effet, plusieurs techniques de fouilles de textes font appel à des méthodes qui ont été utilisées tout d'abord en fouilles de données.
31. Marti HEARST, « What is Text Mining? », 17 octobre 2003, accessible sur <http://people.ischool.berkeley.edu/~hearst/text-mining.html> (traduction libre de l'auteure).
32. Il faut mentionner que ce sont des étapes communes à plusieurs techniques en traitement du langage naturel, mais certains algorithmes peuvent comporter des

corpus ; (3) l'annotation ; (4) l'entraînement du modèle ; et (5) la création du modèle entraîné[33].

La première étape, soit la compilation du corpus nécessaire à la création du modèle, requiert la collecte de ressources linguistiques pertinentes pour l'objectif identifié. Celles-ci doivent posséder des caractéristiques analogues aux données qui seront utilisées lorsque le modèle sera implanté dans des conditions réelles. Par exemple, la langue devrait être la même et si les données sont annotées, les textes servant à l'entraînement du modèle devraient l'être également. Les textes formant ce corpus peuvent provenir de différentes sources, comme des sites Web, des publications, des journaux, ou des articles, académiques ou autres.

La deuxième étape, le prétraitement, consiste en la conversion des textes dans un format de fichier pouvant être traité par l'algorithme. Par exemple, les textes composant les jeux de données sont généralement dans des formats qui ne peuvent être utilisés par les algorithmes (des formats communs étant les PDF ou les fichiers HTML) et doivent donc être transformés en format approprié, comme le *plain text*. Également, certains éléments de ces textes pourront être supprimés, comme les tableaux ou les images, aux fins de l'analyse.

La troisième étape, l'annotation, permet d'identifier des caractéristiques présentes dans le texte en fonction de la méthode utilisée et des catégories choisies (morphologiques, grammaticales, syntaxiques, etc.). Les annotations pourront indiquer, par exemple, les verbes, les noms, ou les adverbes (les classes de mots), mais également d'autres catégories comme des étiquettes identifiant des sentiments, par exemple à des fins de classification ou d'analyse sémantique.

Une fois ces trois étapes complétées, la quatrième étape, l'entraînement, peut avoir lieu. Les algorithmes peuvent ainsi être mis en marche : ils analysent le corpus, donnant comme résultats des informations probabilistes, grammaticales, statistiques et syntaxiques, lesquelles sont sauvegardées dans un fichier.

Finalement, la dernière et cinquième étape est la création du fichier constituant le modèle entraîné. Celui-ci consiste, de façon

étapes supplémentaires. Nous nous restreignons ici à ce flux des travaux à cinq étapes à des fins d'exemple.

33. Thomas MARGONI, « Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI? », *CREATe Working Paper*, University of Glasgow, décembre 2018, p. 3.

concrète, en un outil logiciel ayant extrait les connaissances provenant des données qui lui ont été fournies et qui peut être utilisé afin d'accomplir des tâches spécifiques, comme de la prédiction, sur les données réelles.

Aux fins de l'analyse juridique touchant au droit d'auteur, l'identification et la compréhension de ces étapes est cruciale afin de déterminer si ce processus enfreint les principes reliés à la reproduction.

2. LE DROIT D'AUTEUR ENTOURANT LE TRAITEMENT ET LA COMPRÉHENSION DU LANGAGE NATUREL

La présente section vise à explorer le statut en droit d'auteur des ressources utilisées pour les activités d'apprentissage automatique, spécifiquement du traitement et de la compréhension du langage naturel. Cette description permettra par la suite de procéder à l'analyse des conséquences juridiques des étapes requises pour l'entraînement des algorithmes. Les balises du droit de reproduction prévu par la Loi sont ensuite décrites brièvement afin de délimiter les contours permis et les zones grises.

2.1 Les textes et les œuvres littéraires

Il est important de s'attarder au statut juridique des textes dans le contexte du traitement du langage naturel, car ceux-ci sont les ingrédients clés permettant d'extraire de nouvelles connaissances à l'aide des algorithmes. En effet, chaque base de données qui sera utilisée pour entraîner les algorithmes est composée d'un nombre important de textes uniques provenant de différentes sources aussi variées que des pages Web, des articles de journaux ou des publications académiques et parfois même des publications provenant des médias sociaux. Ces ressources ont généralement le potentiel de faire l'objet de la protection prévue par la Loi, en tant qu'œuvres littéraires[34], sous réserve de respecter les exigences de celle-ci, dont l'originalité.

La notion d'originalité n'est pas définie dans la Loi. Cependant, comme la Cour suprême le rappelle dans sa décision *CCH*[35], pour qu'une œuvre soit considérée comme originale, elle doit émaner de l'auteur, ne pas constituer une copie d'une autre œuvre et résulter de

34. *Loi sur le droit d'auteur*, *supra*, note 1, art. 2.
35. *CCH*, *supra*, note 14, par. 14.

l'exercice non négligeable du talent et du jugement[36]. De plus, ce droit d'auteur s'applique uniquement à l'élément expressif de l'œuvre (sa forme) et non à l'idée sous-jacente : l'œuvre doit donc être concrétisée de manière relativement permanente, ou fixée[37]. Intrinsèquement une question de faits, l'analyse de l'originalité doit donc être faite de façon qualitative et globale[38]. L'unicité n'emporte par ailleurs pas la conclusion automatique qu'une œuvre est originale[39] : deux auteurs pourraient avoir développé de façon indépendante des œuvres similaires, voire identiques, et être originales[40]. Il est à noter, par ailleurs, qu'une compilation peut être jugée originale même si elle porte sur des éléments qui, eux-mêmes, ne le sont pas[41]. De surcroît, en matière de logiciel, les choix arbitraires d'un auteur peuvent être considérés comme créatifs, mais ne pas nécessairement avoir exigé l'exercice de suffisamment de jugement pour mériter la protection de la Loi[42]. C'est le cas notamment où l'agencement de certaines informations est dicté par la réglementation applicable, le médium employé[43], ou encore des considérations fonctionnelles[44]. Il faut cependant mentionner que, malgré ces éléments de précisions apportés par l'arrêt *CCH*, la notion d'originalité, plus particulièrement celle de « talent et de jugement » demeure incertaine quant à sa portée et son application, à une époque où les habitudes en lien avec les données changent rapidement : par exemple, on peut se demander si les photographies prises à l'aide de téléphones cellulaires, les gazouillis sur Twitter, les courriels et les messages textes pourraient être considérés comme triviaux et ne requérant pas nécessairement de talent et de jugement[45].

---

36. *Ibid.*, par. 15 et 25.
37. David VAVER, *Intellectual Property Law*, 2e éd., Toronto, Irwin Law, 2011, p. 135.
38. *Cinar Corporation* c. *Robinson*, 2013 CSC 73, par. 33-37.
39. *Cummings* c. *Global Television Network Quebec Ltd.*, 2005 CanLII 17671 (QC C.S.) (permission d'en appeler refusée : 2007 QCCA 338 ; autorisation du pourvoi rejeté : 2007 CanLII 39170 (C.S.C.)).
40. *Programmation Gagnon inc.* c. *Formules d'affaires CCL inc.*, 2001 CanLII 25191 (QC C.S.).
41. *Hutton* c. *Canadian Broadcasting Corp.*, 1992 ABCA 39, par. 3 ; *Harmony Consulting Ltd.* c. *G.A. Foss Transport Ltd.*, 2011 CF 340 (confirmée en Cour fédérale d'appel sur la question de l'existence du droit d'auteur : 2012 FCA 226).
42. Xavier BEAUCHAMP-TREMBLAY, « Mise à jour en droit d'auteur sur les logiciels à la lumière des récents débats américains (nommément l'affaire *Oracle* c. *Google*) et européens sur l'interopérabilité) », dans S.F.C.B.Q., vol. 389, *Développements récents en droit de la propriété intellectuelle*, Montréal, Éditions Yvon Blais, 2014, p. 182.
43. *Bonnette* c. *Dominion Blueline Inc.*, 2005 QCCA 342, par. 38. Dans cette affaire, il est à noter que bien que l'œuvre en question ne consistait pas en un logiciel, il s'agissait d'une œuvre fonctionnelle pouvant y être assimilée (soit des livres de paie comportant une façon d'organiser des relevés de paie de façon systématique), et donc pertinente pour l'analyse aux présentes.
44. *Delrina Corp.* c. *Triolet Systems Inc.*, 2002 CanLII 11389 (ON C.A.), par. 29.
45. D. VAVER, *supra*, note 37, p. 101.

2.2 La protection des bases de données

Au Canada, les bases de données sont protégées en droit d'auteur en tant que « compilation »[46], laquelle se définit comme suit :

> Les œuvres résultant du choix ou de l'arrangement de tout ou partie d'œuvres littéraires, dramatiques, musicales ou artistiques ou de données. (Nos soulignés)

Afin de bénéficier de la protection de la Loi, ces compilations doivent remplir les critères de celle-ci[47]. L'arrêt phare en cette matière demeure *CCH*, lequel a déterminé que cette protection ne s'applique qu'à la compilation en elle-même et non aux composantes individuelles de celle-ci[48]. Dans ce dossier, les décisions judiciaires publiées dans un recueil ont été jugées originales par la Cour, car les auteurs de celui-ci « ont agencé de façon particulière le résumé jurisprudentiel, les mots-clés, l'intitulé répertorié, les renseignements relatifs aux motifs du jugement (les sommaires) et les motifs de la décision »[49].

Il est à noter que la protection conférée par la Loi est la seule forme de protection disponible pour les bases de données au Canada. En effet, le Canada n'accorde pas de droit distinct aux bases de données autre que celui prévu par le droit d'auteur, contrairement à certaines juridictions, dont l'Europe, qui accordent un droit de protection *sui generis* à certaines bases de données[50]. Ce droit permet une protection unique en son genre, qui n'est pas fondée sur l'originalité, mais qui vise à compenser l'investissement substantiel effectué, d'un point de vue qualitatif ou quantitatif, afin d'obtenir ladite compilation[51]. Cet investissement récompense donc les efforts dans l'obtention, la vérification ou la présentation des données. Le droit est ainsi généralement accordé au fabricant de la base de données, c'est-à-dire celui ayant fait l'investissement financier. Cette protection permet au détenteur du droit d'empêcher les extractions et la réutilisation de la totalité ou d'une partie substantielle de la base de données protégée[52].

La jurisprudence récente révèle que le domaine des données, plus particulièrement leur création et leur utilisation, fait l'objet

46. *Loi sur le droit d'auteur*, *supra*, note 1.
47. *CCH*, *supra*, note 14.
48. *Ibid.*, par. 33.
49. *Ibid.*, par. 34.
50. *Directive 96/9/CE concernant la protection juridique des bases de données*, OJL 77, 27.3 1993.
51. *Ibid.*, art. 1.
52. *Ibid.*, art. 7.

d'opinions divergentes quant à l'interprétation de la portée de l'exercice du talent et du jugement. Par exemple, la décision albertaine *Geophysical Service Incorporated* c. *EnCana Corporation*[53] (« GSI ») a déterminé que les données sismiques, dans leur forme brute et traitée, constituaient une œuvre susceptible d'être protégée par droit d'auteur, puisqu'elles avaient été générées par l'exercice de talent et de jugement, avec l'aide d'ordinateurs. La Cour a ainsi reconnu dans cette affaire que les « coupes sismiques », soit des représentations graphiques de données traitées, affichées et interprétées par des géophysiciens professionnels, exigeaient le choix ou l'organisation des données sous-jacentes[54]. De plus, les membres de l'équipe avaient dû utiliser leur jugement afin de créer ces données brutes en choisissant les bons emplacements, le positionnement des instruments, la préparation des appareils recueillant les données, etc.

Comme nous le verrons plus loin, cette conclusion dans *GSI* quant aux données brutes pourrait avoir une incidence importante en IA, car, au-delà de la compilation des données, la composition des items individuels de ces jeux de données (particulièrement quant aux données transformées, qui peuvent faire l'objet d'une certaine analyse et habileté) pourrait bénéficier de la protection du droit d'auteur et, ainsi, en dériver une certaine valeur commerciale exploitable par ses propriétaires.

Cependant, dans l'arrêt *Toronto Real Estate Board* c. *Commissaire de la concurrence*[55] (« TREB »), la Cour d'appel fédérale a conclu que la base de données Multiple Listing Service[MD] ne pouvait faire l'objet de la protection du droit d'auteur. Dans ce dossier, les bases de données étaient compilées à partir de données factuelles, assemblées par les courtiers REALTOR[R] et entrées dans la base de données manuellement, entrainant une apparition quasi instantanée des données dans celle-ci. La Cour a donc conclu que cette compilation ne correspondait pas au critère d'originalité prévu par la Loi, car cette saisie ne faisait pas l'objet de l'exercice du talent et du jugement.

Force est d'admettre que cette conclusion remet en question l'interprétation qui pourra être faite des compilations de données réalisées à l'aide de l'IA. En effet, la question de juger si une œuvre résulte d'une entreprise purement mécanique est contextuelle et

53. 2016 ABQB 230, confirmée en partie en appel, notamment sur la question relative au droit d'auteur (2017 ABCA 125); demande d'autorisation en Cour suprême rejetée (2017 CanLII 80435 (CSC)) [*GSI*].
54. 2016 ABQB 230, par. 79-81.
55. 2017 CAF 236 [*TREB*].

basée largement sur des éléments factuels. En l'espèce, dans *TREB*, la Cour fédérale semble justifier sa décision principalement sur la preuve limitée présentée devant la Cour et sur le fait que les données de la base de données étaient générées « presque instantanément » suivant les normes de l'industrie. Or, dans le cadre du prétraitement des données servant à entraîner les algorithmes d'IA, les paramètres choisis pour cette étape pourraient, dans certains cas, être considérés comme originaux, bien qu'ils puissent être circonscrits notamment par certaines règles de l'art (similaires aux normes de l'industrie mentionnées dans *TREB*) en matière de traitement du langage naturel.

2.3 Les adaptations et les autres œuvres dérivées

L'importance de la notion d'adaptation provient de la question relative aux données transformées, soit le résultat généré suivant l'entraînement d'un algorithme. Il est à noter que le concept d'œuvre dérivée n'est pas prévu expressément dans la Loi. Par contre, la Cour suprême a reconnu, dans l'arrêt *Théberge*[56], qu'il existe en droit canadien une protection limitée pour ce type d'œuvre. Le tribunal a jugé dans cette affaire que « les mots "produire ou reproduire [...] l'œuvre, sous une forme matérielle quelconque" figurant au par. 3(1) confèrent aux artistes et aux auteurs le droit exclusif de contrôler la préparation des œuvres dérivées »[57].

On peut en déduire qu'au Canada, la reproduction d'une œuvre dérivée sera donc considérée comme une reproduction illicite, à moins de bénéficier d'une exception prévue par la Loi ou de constituer une œuvre originale distincte[58].

Il est important de mentionner que ces droits limités ne peuvent être comparés à la protection explicite plus étendue prévue dans d'autres juridictions, telles que les états-Unis[59]. En effet, le droit

56. *Théberge*, *supra*, note 14, par. 71.
57. *Ibid.*, par. 73.
58. À cet effet, une décision en Nouvelle-Écosse a distingué des « plans concepts » de plans architecturaux en raison des différences marquées entre les plans concepts et les plans finaux, indiquant que les seules similarités résultaient des contraintes imposées par le client en matière de design. Voir *MacNutt* c. *Acadia University*, 2016 NSSC 160 (appel rejeté en Cour d'appel de la Nouvelle-Écosse (2017 NSCA 57) et demande d'autorisation en appel à la Cour suprême également rejetée (2018 CanLII 33051 (C.S.C.)).
59. La loi américaine, contrairement au droit canadien, incorpore expressément une définition d'œuvre dérivée, laquelle est considérée comme une nouvelle œuvre pouvant faire l'objet de la protection du droit d'auteur. Le titulaire du droit d'auteur a donc le droit additionnel (en plus de la reproduction) de contrôler la création d'une telle œuvre : voir 17 U.S.C. §101.

américain reconnait une protection sur « toute autre forme dans laquelle une œuvre peut être refaçonnée, transformée ou adaptée »[60].

2.4 Le droit de reproduction et ses paramètres

Après avoir décrit le statut juridique des ressources utilisées par les algorithmes, il faut s'attarder à la description des paramètres du droit de reproduction. Comme mentionné plus haut, le droit exclusif de reproduire une œuvre et de contrôler les reproductions faites de cette dernière est au cœur de l'application du droit d'auteur canadien[61]. Il s'ensuit donc que sous réserve d'une exception expressément prévue par la Loi, l'auteur d'une œuvre protégée peut exercer son droit et faire cesser les violations en conséquence[62].

Ces exceptions sont exhaustives et visent des fins spécifiques[63], ainsi que des entités particulières, comme les établissements d'enseignement[64] et les bibliothèques, les musées et les services d'archives[65].

La Loi prévoit également des exceptions reconnaissant de façon explicite la nature des technologies actuelles : c'est le cas notamment de l'exception concernant les reproductions temporaires pour processus technologiques[66]. Trois conditions doivent être remplies afin de déterminer si l'exception s'applique : (1) la reproduction est un élément essentiel d'un processus technologique ; (2) elle a pour seul but de faciliter une utilisation qui ne constitue pas une violation du droit d'auteur ; et (3) elle n'existe que pour la durée du processus technologique. Le caractère « essentiel » d'une reproduction du premier critère pourrait être matière à débat, mais on peut présumer que l'intention est de favoriser l'efficience des processus technologiques, plutôt que d'interpréter le terme de façon indûment stricte[67]. Par contre, nous verrons dans la section 3 que la troisième condition peut s'avérer difficile à respecter dans le cadre des processus d'apprentissage automatique.

---

60. *Ibid*.
61. *Loi sur le droit d'auteur*, *supra*, note 1.
62. *Ibid.*, art. 27(1)-27.1.
63. *Ibid.*, art. 29-29.3 ; 30.6-30.7 (relatifs aux programmes d'ordinateurs).
64. *Ibid.*, art. 29.4-30.04.
65. *Ibid.*, art. 30.1-30.5.
66. *Ibid.*, art. 30.71.
67. Cameron HUTCHISON, *Digital Copyright Law*, Toronto, Irwin Law, 2016, p. 169 (sur le fait que d'interpréter le mot « essentiel » comme un synonyme « indispensable » reviendrait à décourager ce type de processus technologique).

Malgré cette exception, le droit de reproduction peine à entrer dans l'ère moderne. En effet, contrairement à d'autres droits qui ont dû forcément évoluer avec l'avènement de nouvelles technologies, le droit de reproduction demeure peu flexible et ne contient que des exceptions spécifiques (à l'exception de l'utilisation équitable). Certains auteurs suggèrent que cette tension peut s'expliquer par la tâche ardue de distinguer qui sont les utilisateurs du droit d'auteur, c'est-à-dire les « consommateurs » des œuvres, soit les individus, mais aussi les utilisateurs techniques ou commerciaux, soit les intermédiaires, qui rendent disponibles ces œuvres et les auteurs[68]. Le difficile exercice auquel doivent se livrer les tribunaux afin de trouver un équilibre entre les droits des différentes parties intéressées, particulièrement dans un contexte où la jurisprudence envoie des signaux mixtes quant à l'identité et les intérêts de ces utilisateurs, rend difficile la gestion de risques pour ceux-ci[69].

Cette tension est illustrée de façon particulièrement accrue dans le cadre des processus technologiques inhérents à l'IA, où des techniques faisant maintenant partie des règles de l'art en apprentissage automatique mettent en lumière la rigidité des règles relatives à la reproduction et démontrent clairement l'incompatibilité avec l'état de la recherche actuelle. De plus, l'absence de reconnaissance explicite dans la Loi et la jurisprudence des intérêts de nouveaux joueurs, comme les entreprises privées œuvrant en recherche, crée plusieurs zones grises. Des modifications législatives repensant ce droit à l'ère de la révolution technologique actuelle propulsée par l'IA semblent donc nécessaires. Finalement, ainsi qu'il sera décrit plus amplement dans la prochaine section, l'incertitude quant à l'application des exceptions contribue à augmenter les risques pour les entreprises innovantes.

2.5 L'exception de l'utilisation équitable

Comme mentionné ci-dessus, la structure de la Loi permet l'usage d'une œuvre protégée lorsque cet usage fait l'objet d'une exception explicite. Cependant, la Loi prévoit également une exception plus large et contextuelle, soit celle de l'utilisation équitable.

La Cour suprême, dans *CCH*, définit cette exception, à l'instar des autres exceptions, non pas comme une simple défense à la violation du droit d'auteur, mais plutôt comme un droit des utilisa-

68. Ysolde GENDREAU, « Walking the Copyright Tight Rope », dans *Research Handbook on Copyright Law: Second Edition*, Éditions Edward Elgar, 2017, p. 28.
69. *Ibid.*, p. 29.

teurs[70]. À ce titre, la Cour réitère que cette exception ne doit pas être interprétée restrictivement, afin de maintenir un équilibre entre les titulaires du droit d'auteur et les utilisateurs[71], mais qu'il faut plutôt favoriser l'approche décrite ci-dessous. De plus, pour analyser l'utilisation de l'œuvre, il faut se situer du point de vue de l'utilisateur et non du titulaire du droit d'auteur[72].

Les tribunaux canadiens ont adopté, depuis *CCH*, une approche en deux volets afin de déterminer si une utilisation d'une œuvre protégée est équitable[73]. Le premier volet vise à analyser si l'utilisation est autorisée pour une des fins permises, soit la satire, la parodie, l'étude privée et l'éducation[74]. La deuxième étape consiste ensuite en l'évaluation du caractère équitable, basé sur les six facteurs non exhaustifs suivants :

> (1) le but de l'utilisation ;
>
> La fin réellement poursuivie dans l'utilisation de l'œuvre doit être une des fins permises par la Loi. Les tribunaux doivent, dans l'évaluation de ce premier critère, tenter de déterminer de façon objective le but ou le motif réel de l'utilisation en question[75].
>
> (2) la nature de l'utilisation ;
>
> Le critère de la nature de l'utilisation, quant à lui, vise la manière dont l'œuvre a été utilisée. La Cour, dans *CCH*, men-

70. *CCH*, *supra*, note 14, par. 48.
71. Cette analyse a été récemment confirmée dans la décision *Canadian Copyright Licensing Agency* c. *Université York* (2017 CF 669 ; confirmée en Cour d'appel fédérale), où les lignes directrices établies par l'Université comprenait des seuils fixes permettant de tracer la ligne entre une utilisation équitable et inéquitable d'œuvres protégées : la Cour a déterminé que ces seuils étaient arbitraires et non fondés en principe, confirmant ainsi l'importance de maintenir une approche flexible et contextuelle lors de cette analyse.
72. *SOCAN* c. *Bell Canada*, 2012 CSC 36, par. 30 et 34 [*Bell*].
73. Il est à noter que certains auteurs remettent en question la notion selon laquelle un résultat négatif suivant l'analyse du premier volet ferait échec au test, arguant qu'il ne peut s'agir de l'intention du législateur d'avoir restreint indûment l'interprétation de cette exception. Voir Ariel KATZ, « Debunking the Fair Use vs. Fair Dealing Myth: Have We Had Fair Use All Along? », dans l'ouvrage à paraître *Comparative Aspects of Limitations and Exceptions in Copyright Law* (Shyam BALGANESH, Wee Loon NG-LOY, Haochen SUN (dir.), 2018).
74. *Loi sur le droit d'auteur*, *supra*, note 1, art. 29.
75. C. HUTCHISON, *supra*, note 67, p. 149 (sur la distinction importante entre ce premier critère et le deuxième, soit la nature de l'utilisation, en ce que l'on recherche, d'une part, le véritable objectif, mais également le but plus spécifique de l'utilisateur).

tionne que la distribution de multiples copies tend à démontrer une utilisation inéquitable, tandis que la distribution d'une seule copie utilisée à des fins légitimes en particulier, ou encore le fait que la copie ait été détruite, amènerait aisément à la conclusion que celle-ci était équitable[76]. Il est à noter que la dissémination massive sur Internet de copies ne serait pas nécessairement un indicateur automatique d'une utilisation inéquitable, à condition toutefois que le but de celle-ci soit équitable[77]. Par ailleurs, dans une autre décision, le tribunal clarifie que l'absence de consentement à l'activité en question n'indique pas la fin de l'analyse, mais plutôt le début de l'analyse sur la nature de celle-ci[78].

(3) l'ampleur de l'utilisation (l'ampleur de la reproduction) ;

Ce troisième critère fait référence à l'ampleur de l'utilisation de façon à la fois quantitative et qualitative. De plus, la fin poursuivie informe la conclusion quant au caractère équitable. En effet, une œuvre pourrait être utilisée dans son entièreté en raison de la nature de celle-ci : c'est le cas des photographies[79].

(4) les solutions de rechange à l'utilisation ;

L'existence d'une autre option que l'utilisation de l'œuvre protégée peut également influer sur l'évaluation du caractère équitable ou inéquitable de celle-ci[80]. Le caractère raisonnablement nécessaire de l'utilisation de l'œuvre pourra également avoir une influence. Cependant, une alternative trop onéreuse en pratique, telle que l'assemblage des informations de façon manuelle alors qu'un moyen automatique est disponible et en lien avec le but de l'utilisation, ne pourrait pas être raisonnablement considérée comme une véritable solution de rechange[81].

---

76. *CCH*, *supra*, note 14, par. 55.
77. C. HUTCHISON, *supra*, note 67, p. 150 (en référence au commentaire dans *Bell*, *supra*, note 72, sur l'importance de prendre en considération le principe de la neutralité technologique afin d'éviter de désavantager l'Internet par rapport à d'autres médias).
78. *Century 21 Canada Limited Partnership* c. *Rogers Communications Inc.*, 2011 BCSC 1196, par. 252 [*Century 21*].
79. *CCH*, *supra*, note 14, par. 56.
80. *Ibid.*, par. 57.
81. *Century 21*, *supra*, note 78, par. 262 ; voir également C. HUTCHISON, *supra*, note 67, p. 154-155.

(5) la nature de l'œuvre ; et

Bien que non décisif, le cinquième critère, c'est-à-dire le fait que l'œuvre soit non publiée ou confidentielle, peut également aider à la détermination du caractère équitable[82]. Ce facteur prend en considération un facteur extrinsèque au droit d'auteur, soit la confidentialité, et le met en relief avec l'objectif de la Loi, soit de favoriser la création et de disséminer les œuvres. Dans ce cas, le fait de rendre publique une œuvre tendrait à pencher vers une conclusion que l'utilisation est équitable.

(6) l'effet de l'utilisation sur l'œuvre.

Ce sixième et dernier critère fait référence à l'effet que pourrait avoir l'utilisation sur l'œuvre, tel que la concurrence sur celle-ci[83]. Une utilisation jugée complémentaire à celle de l'auteur serait donc vue comme plus équitable[84].

3. L'APPLICATION DE LA *LOI SUR L'ENTRAÎNEMENT DES MODÈLES DE TRAITEMENT ET DE COMPRÉHENSION DU LANGAGE NATUREL*

3.1 Les reproductions temporaires et permanentes

Comme il a été décrit plus tôt, dans le cas de l'IA, une incertitude réside dans le fait qu'à l'étape du prétraitement, la création d'une copie temporaire est nécessaire à l'exécution des tâches subséquentes. La question qui se pose alors est la suivante : quel est le statut de cette copie ? Respecte-t-elle les exigences de la Loi ou est-elle couverte par l'exception de l'utilisation équitable ?[85]

Il s'avère que la réponse à cette question est loin d'être claire. En effet, par exemple, dans un contexte commercial par une entreprise privée, la plupart des exceptions sont d'emblée exclues. Tout d'abord, l'entreprise privée ne bénéficie d'aucune exception basée strictement sur son statut, par opposition aux établissements d'enseignement ou encore aux bibliothèques. Également, les exceptions permises concernant les logiciels ne trouvent pas application dans l'entraîne-

82. *CCH*, *supra*, note 14, par. 58.
83. *Ibid.*, par. 59.
84. *Bell*, *supra*, note 72, par. 48 ; *Century 21*, *supra*, note 78, par. 275 (*Century 21* ayant été décidée avant *Bell*, il semble que la Cour ait changé d'avis quant à l'application de ce critère).
85. *Loi sur le droit d'auteur*, *supra*, note 1, art. 29 et s.

ment des algorithmes, car ils se limitent à des processus techniques routiniers comme la création de copies pour la sauvegarde[86], à des fins de compatibilité[87] ou encore d'interopérabilité[88].

Qu'en est-il de l'exception visant les reproductions temporaires pour processus technologiques ? On peut naturellement envisager au premier abord cette exception comme une solution potentiellement « neutre » d'un point de vue technologique. Toutefois, une lecture plus détaillée de l'article 30.71 nous amène à la conclusion que bien que deux des conditions puissent tenir la route, la troisième condition, soit la durée de l'existence de la copie temporaire, semble créer un problème direct d'application dans le cadre de l'entraînement d'un algorithme.

Bien qu'il existe peu de jurisprudence sur le sujet, la portée de cet article a été analysée dans une décision de la Commission du droit d'auteur du Canada (« Commission »)[89]. La Commission a indiqué dans cette affaire que sa lecture de ce critère « donnait à penser que la destruction d'une telle copie est vraisemblablement automatique dans le cadre de la technologie utilisée »[90]. La Commission a poursuivi son raisonnement en ajoutant que lorsque la Loi indiquait que les copies pouvaient être conservées pendant une certaine période, et qu'un geste concret devait intervenir pour leur destruction, cette dernière était expressément incluse comme obligation.

Il est à noter que l'exception canadienne n'est pas unique : une exception européenne similaire existe, à quelques différences près[91]. La différence majeure consiste néanmoins dans le remplacement du critère de la durée de la reproduction par un troisième critère visant la nature de l'usage, soit l'absence de valeur économique élevée

86. *Ibid.*, art. 30.6(b).
87. *Ibid.*, art. 30.6(a).
88. *Ibid.*, art. 30.61.
89. COMMISSION DU DROIT D'AUTEUR DU CANADA, *Tarif des redevances à percevoir par SOCAN, Ré:Sonne, CSI, Connect/SOPROQ et Artisti à l'égard des stations de radio commerciale*, 2016, accessible sur : <https://cb-cda.gc.ca/decisions/2016/DEC-2016-04-21.pdf>.
90. *Ibid.*, par. 181.
91. *Directive 2001/29/CE du Parlement européen et du Conseil du 22 mai 2001 sur l'harmonisation de certains aspects du droit d'auteur et des droits voisins dans la société de l'information*, art. 5(1). L'article réfère aux « actes de reproduction provisoires visés à l'article 2, qui sont transitoires ou accessoires et constituent une partie intégrante et essentielle d'un procédé technique et dont l'unique finalité est de permettre a) une transmission dans un réseau entre tiers par un intermédiaire, ou b) une utilisation licite d'une œuvre ou d'un objet protégé, et qui n'ont pas de signification économique indépendante […] ».

indépendante résultant d'un tel usage. L'interprétation de la portée de cette disposition a été clarifiée dans la décision *Infopaq International* c. *Danske Dagblades Forening*[92] (« Infopaq »). Dans cette affaire, la Cour devait se prononcer sur l'application de l'exception relative aux copies transitoires et accessoires à un processus similaire à celui utilisé pour le traitement du langage naturel. Comme nous le verrons ci-dessous au paragraphe 3.2, de nombreuses similitudes amènent un éclairage intéressant sur l'application de l'exception au contexte spécifique du traitement du langage naturel.

Bien que certains commentateurs avancent que la création de telles copies transitoires pourrait être exclue de l'application de la Loi[93], il demeure tout de même qu'en l'absence de clarté sur la question, il faut examiner l'application de la seule exception qui subsiste véritablement, soit celle de l'utilisation équitable. L'analyse en deux volets prévue par *CCH* démontre pourtant qu'une certaine incertitude demeure quant à son application. En effet, en ce qui concerne le premier volet, les fins permises énumérées à l'article 29 de la Loi étant exhaustives, seule la recherche peut être envisageable pour le traitement du langage naturel.

En excluant le contexte classique de la recherche menée exclusivement dans un établissement d'enseignement, on peut s'interroger dans les cas où la recherche est effectuée dans un contexte avec un but commercial, par exemple lorsqu'elle est le fruit d'efforts en entreprise privée, ou dans le cadre d'un projet de recherche public-privé. D'une part, la jurisprudence reconnaît que la recherche à des fins commerciales peut être permise dans certaines circonstances[94]. En effet, la Cour suprême, dans *CCH*, a rappelé qu'il fallait « interpréter le mot "recherche" de manière large afin que les droits des utilisateurs ne soient pas indûment restreints, et que la recherche ne se limite pas à celle effectuée dans un contexte non commercial ou privé »[95].

Cela étant dit, dans le cas de la recherche commerciale, malgré tout, un flou demeure. À cet effet, la Cour suprême, dans *Alberta (Education)* c. *Canadian Copyright Licensing Agency (Access Copyright)*[96], distingue le cas à l'étude, où l'enseignant et l'étudiant sont considérés comme étant en symbiose aux fins de l'analyse du but de l'utilisation, d'autres décisions comportant des trames factuelles similaires où

92. C-5/08 [2009] E.C.R. I-06569 ; confirmé en appel en 2012, C-302/10 [*Infopaq*].
93. C. HUTCHISON, *supra*, note 67, p. 156.
94. *CCH*, *supra*, note 14, par. 51.
95. *Ibid*.
96. 2012 CSC 37.

les intéressés étaient de toute évidence animés par un motif inavoué, tel un motif commercial. S'appuyant sur les fins permises que sont la « recherche ou l'étude privée », ils tentaient en fait de s'attribuer les fins de leurs clients ou des étudiants afin d'échapper à des allégations de droit d'auteur[97].

Ces décisions, bien que distinguées par la Cour dans ce dossier, démontrent que l'identité de l'utilisateur, combinée au but de celui-ci, peut amener des conclusions contradictoires malgré des faits similaires.

Une analyse du deuxième volet, soit la détermination du caractère équitable, renforce l'argument de l'inadéquation du langage prévu par l'exception et celui de la réalité technologique.

Tout d'abord, le but de l'utilisation est un des facteurs comportant le plus d'ambiguïtés quant à son arrimage avec l'intention législative de la Loi[98]. Ici, la fin réellement poursuivie dans l'utilisation de l'œuvre peut poser certains obstacles : quel est le motif réel d'une entreprise lorsqu'elle entraîne un algorithme ? Peut-on parler de recherche à des fins internes si celui-ci est intégré à une date ultérieure dans un produit commercial ou dans un projet pour un client ? Qu'en est-il des projets conjoints entre un établissement d'enseignement et une entreprise privée ? Ces exemples ne sont que quelques cas de figure qui mettent en lumière la nécessité d'avoir une exception suffisamment flexible pour permettre l'émergence de nouveaux modèles d'affaires.

En second lieu, la nature de l'utilisation peut également être considérée comme étant en défaveur d'une conclusion d'utilisation équitable. En effet, on pourrait prétendre qu'il est facétieux d'arguer que la compréhension du langage naturel n'a qu'une fin « technique »[99]. Une telle prétention nierait la valeur sémantique et intrinsèque de tout texte. Par contre, d'autres pourraient avancer que l'utilisation, particulièrement à des buts d'analyse non sémantique,

97. *Ibid.*, par. 20-21.
98. Comme le mentionne la Cour dans *Théberge*, *supra*, note 14, par. 10 : « La Loi est généralement présentée comme établissant un équilibre entre d'une part, la promotion, dans l'intérêt du public, de la création et de la diffusion des œuvres artistiques et intellectuelles et, d'autre part, l'obtention d'une juste récompense pour son créateur […] ».
99. En effet, les tâches de compréhension du langage naturel ont pour but principal d'extraire le sens des textes analysés : on peut penser notamment à des tâches concrètes comme la production de résumés, l'analyse de sentiments, ou encore les agents conversationnels.

comme dans certains cas en traitement du langage naturel[100], reviendrait à un processus mécanique qui ne devrait pas être considéré comme inéquitable.

Quant à l'ampleur de l'utilisation ou de la reproduction, ce facteur semble plutôt neutre si l'on assimile cette reproduction à celle des photos. En effet, l'extraction d'informations d'un texte ne peut se faire qu'en utilisant l'entièreté de celui-ci[101]. De plus, le texte n'est utilisé présumément qu'une seule fois lors du processus d'entraînement ou, tout au plus, un nombre limité de fois, ce qui renforce l'argument de la neutralité de l'analyse de ce critère quant à l'aspect quantitatif de l'usage.

Si on prend en considération les solutions de rechange à l'utilisation des œuvres protégées, il faut constater les limites inhérentes au niveau de maturité des ressources requises par la technologie actuelle. Effectivement, le fait que bon nombre de bases de données accessibles publiquement soient tout d'abord compilées à des fins de recherche académique et que celles-ci ne trouvent pas nécessairement leur équivalent dans les bases de données commerciales constitue la preuve, dans plusieurs cas, qu'il n'existe pas réellement de solutions de rechange à l'utilisation des œuvres.

Néanmoins, on peut raisonnablement se questionner sur le contenu de ces bases de données, qui pourrait, selon le cas, être modifié afin d'utiliser d'autres œuvres, possiblement provenant du domaine public[102]. En contrepartie, le fardeau imposé en termes de temps et de ressources nécessaires afin de substituer une base de données existante ou de créer une nouvelle base de données exempte d'œuvres protégées par le droit d'auteur, serait l'argument automatiquement soulevé par les utilisateurs[103] : quel serait le niveau requis afin de

---

100. Les tâches de la fouille de textes, comme la classification, ne permettent pas d'effectuer une analyse sémantique en elle-même, mais bien d'extraire une certaine valeur dans un volume important de textes.
101. Cet argument a d'ailleurs été accepté par des tribunaux américains : voir *Authors Guild, Inc.* c. *HathiTrust*, F. 3d (2nd Cir. 2014), p. 98 [*HathiTrust*] et *Authors Guild* c. *Google*, No. 13-4829-cv (2nd Cir. October 16, 2015), p. 21.
102. Sur ce sujet, voir Megan SENSENEY, Beth NAMACHCHIVAYA, Eleanor DICKSON *et al.*, « Data Mining Research with In-copyright and Use-limited Text Datasets: Preliminary Findings from a Systematic Literature Review and Stakeholder Interviews », article publié dans le cadre de la conférence IDCC18, 21 janvier 2018 [*IDCC18*].
103. COMMISSION DU DROIT D'AUTEUR DU CANADA, *supra*, note 89, par. 133. La Commission, dans sa décision, reconnaît cependant qu'elle ne peut pas « imposer aux diffuseurs un fardeau de preuve si exigeant qu'il leur serait impossible à l'avenir de démontrer qu'ils satisfont aux exigences d'une exception.

respecter le caractère équitable ? On peut également s'interroger sur la valeur d'une telle base de données, laquelle pourrait avoir nécessité lors de sa création de faire des choix compromettant la qualité de celle-ci simplement afin de contourner les obstacles posés par la Loi.

Quant à la nature de l'œuvre, le cinquième critère, il tend à renforcer une interprétation neutre, compte tenu de la nature transitoire de l'usage de l'œuvre et de la fin d'extraction d'informations, ainsi que du caractère d'intérêt public dans certain cas, notamment où la fouille de textes permet un meilleur accès à l'information[104].

Finalement, le dernier critère, soit l'effet de l'utilisation sur l'œuvre, est également moins contentieux, particulièrement dans le cas où l'entraînement de l'algorithme est effectué à des fins internes (par exemple afin de tester un produit ou une approche de résolution d'un problème complexe). L'usage n'est dans ce cas fait non pas pour concurrencer les œuvres originales, mais bien en extraire la valeur informationnelle.

3.2 Les processus de capture d'information et le traitement du langage naturel

L'exception des reproductions temporaires pour processus technologiques n'a pas fait l'objet, depuis son entrée en vigueur en 2012, d'une analyse jurisprudentielle approfondie. En effet, une des seules décisions recensées est celle de la Commission mentionnée plus haut, et celle-ci vise une situation distincte, soit la diffusion de musique en continu (*streaming*).

Par contre, la Cour européenne de justice, dans la décision *Infopaq*, a eu l'occasion de traiter d'un dossier dont les éléments factuels se rapprochaient davantage du processus technique du traitement du langage naturel. Dans cette affaire, la Cour devait

---

Dans les instances en matière de tarif, le nombre d'œuvres ou d'autres objets de droit d'auteur à considérer est colossal, et la qualité et le détail des éléments de preuve relatifs à une copie précise ne sont vraisemblablement pas les mêmes que ceux présentés dans le cadre d'une action en contrefaçon. Le fait d'établir un seuil trop exigeant pour démontrer la conformité des utilisations futures aux exigences d'une exception aurait pour effet, à toutes fins utiles, de rendre théoriques les exceptions créées par le législateur, ce qui serait inéquitable et contraire à l'objet de la Loi ».

104. *Kelly* c. *Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003), p. 819 (c'est l'argument dans cette décision américaine ou la création d'icônes à faible résolution permettant de retrouver des images originales sur les sites Web a été considérée comme améliorant l'accès à l'information).

analyser le modèle d'affaires d'Infopaq, soit des services de veille médiatiques, afin de déterminer s'il y avait lieu de l'exempter d'une contravention au droit d'auteur en raison de l'exception pour copies transitoires[105]. Ces services consistaient de façon sommaire en la compilation, l'extraction, l'indexation et l'impression d'articles de journaux et de mots-clés. Les données afin de produire les rapports étaient obtenues à l'aide d'un processus que la Cour a synthétisé en cinq étapes[106] : (1) l'entrée de données manuelles dans une base de données d'articles de journaux ; (2) la création d'un fichier de format TIFF à partir de sections sélectionnées, scannées, puis transférées à un serveur de reconnaissance optique de caractères ; (3) la transformation de chacune des lettres des fichiers TIFF en un caractère reconnu par l'ordinateur et sauvegardé comme fichier de format texte (le fichier TIFF étant par la suite supprimé) ; (4) le traitement du fichier texte afin d'identifier les mots-clés recherchés, ainsi que la capture des cinq mots précédant et suivant chacun des mots-clés, avant que ce fichier texte soit également effacé ; et finalement, (5) l'impression d'une fiche résumée, comprenant les pages correspondant aux résultats de recherche, ainsi que les courts extraits (c'est-à-dire les extraits de 11 mots consécutifs) de ces pages.

La Cour a déterminé que les quatre premières étapes étaient assujetties à l'exception relative aux copies transitoires, tandis que l'étape cinq était exclue, car une copie permanente d'une œuvre protégée était effectuée. Le tribunal a en effet considéré l'étape de l'impression du rapport, lequel contenait les courts extraits (qui, nous devons le rappeler, étaient de 11 mots), comme une reproduction partielle donnant lieu à une violation du droit d'auteur. Cette reproduction partielle, bien que très courte d'un point de vue quantitatif, pouvait tout de même permettre de détecter la création intellectuelle de l'auteur, en d'autres mots le seuil d'originalité nécessaire afin de permettre la protection prévue par le droit d'auteur.

Lorsqu'on applique ces principes au traitement du langage naturel, on peut manifestement observer plusieurs similitudes. D'une part, la première étape, soit la création de la base de données électronique via la sélection de textes pertinents, s'apparente à l'étape de compilation du corpus décrite plus haut. De plus, la deuxième étape de conversion du fichier TIFF peut être assimilée à celle du prétraitement permettant de convertir les textes en format lisible pour les outils de traitement du langage naturel. La troisième étape, quant à elle, est similaire à celle de l'annotation en ce que l'image est

105. *Infopaq*, *supra*, note 92.
106. *Ibid.*, par. 16-21.

transformée afin de permettre à l'ordinateur de la reconnaître. L'étape quatre, soit l'identification des éléments pertinents, peut aussi être mise en parallèle avec l'entraînement de l'algorithme. Finalement, la dernière étape d'impression du rapport peut s'apparenter à celle de la création et la sauvegarde du modèle entraîné.

On peut ainsi en déduire que le processus d'apprentissage automatique pourrait faire l'objet de l'exception européenne en partie (en excluant la copie permanente de la dernière étape), sous réserve, encore une fois, de respecter le dernier critère d'absence de gain économique indépendant dérivé de l'utilisation de ces copies temporaires. Il semble vraisemblablement que le profit généré serait lié au modèle, plutôt qu'à la copie en elle-même, compte tenu de la clarification de la Cour que les gains d'efficience résultant du processus technologique ne sont pas pris en considération dans l'analyse[107].

Bien que s'appuyant sur le cadre législatif européen, *Infopaq* fournit un éclairage intéressant dont il est possible de s'inspirer dans le cas qui nous occupe. Cette décision donne un aperçu des interprétations possibles à partir de ce type d'exceptions, mais démontre sa faiblesse à résoudre les problèmes posés par le traitement du langage naturel en raison de leur rigidité[108].

3.3 Les licences des bases de données : un obstacle contractuel et terminologique

Dans un autre ordre d'idées, il serait impensable de discuter des obstacles pratiques au traitement du langage naturel sans mentionner les défis supplémentaires auxquels font face les acteurs de l'industrie au point de vue contractuel. En effet, l'accès aux ressources nécessaires afin d'entraîner les algorithmes peut s'avérer une contrainte de taille, et ce, même si le résultat de l'analyse juridique est positif quant à la conformité aux règles entourant le droit d'auteur. Les bases de données peuvent entre autres comporter des problèmes de plusieurs types, dont les modalités restrictives de la licence, mais également une ambiguïté quant aux droits touchant au contenu de celles-ci.

---

107. *Infopaq*, *supra*, note 92, par. 51.
108. Il est impossible de passer sous silence qu'une des raisons de cette rigidité dans le texte de la Loi s'explique en partie par les pressions des lobbys de certaines industries, dont celle de la musique, ainsi que celle d'autres acteurs impliqués dans la protection collective des droits des auteurs, qui exercent une influence importante dans les discussions liées à toute réforme du droit d'auteur et aux modifications législatives subséquentes.

Dans le premier cas, la difficulté réside dans le fait que plusieurs bases de données accessibles publiquement proviennent du milieu universitaire[109]. Lorsqu'une base de données est compilée à des fins de recherche académique, les modalités des bases de données créées à ces fins contiennent de façon générale une interdiction d'usage à des « fins commerciales », sans que la portée de cette expression soit définie. D'un point de vue strictement juridique, comme décrit plus haut, en vertu de *CCH*, la recherche peut être effectuée dans un but commercial et tout de même bénéficier de l'exception de l'utilisation équitable. Cependant, en l'absence de langage contractuel précis[110], dans un contexte où les licences sont très courtes et généralement rédigées comme une arrière-pensée[111], les utilisateurs potentiels de ces bases de données sont incapables de déterminer avec certitude la portée de cette restriction.

De l'autre côté du spectre, certaines bases de données sont exploitées par des entités commerciales. Dans ce contexte, le problème inverse se pose. Il s'agit d'une réalité à laquelle sont confrontés notamment les bibliothèques et les établissements universitaires, où, au contraire, les licences des bases de données nécessaires à leur travail sont trop restreintes[112]. En pratique, deux situations peuvent survenir : (1) la licence prohibe un tel usage, peu importe que celui-ci puisse être considéré comme équitable ou non ; ou (2) la licence est silencieuse sur ce point. Actuellement, peu de licences offertes sur le marché contiennent une telle exception dans les modalités de licence. Le rapport de force inégal entre l'utilisateur (l'université, la bibliothèque, ou même l'entreprise privée) amène souvent l'utilisateur à devoir accepter des conditions moindres que celles souhaitées, simplement afin de garantir l'accès continu à une source d'information précieuse (particulièrement lorsqu'il s'agit de bases de données

109. De nombreuses bases de données utilisées couramment ont été créées tout d'abord dans le cadre de projets de recherche universitaires ou de concours académiques. C'est le cas notamment de ImageNet (accessible sur <http://image-net.org/about-overview>) et de Snap Amazon Reviews (accessible sur <https://snap.stanford.edu/data/web-Amazon.html?source=post_page----------------------------->), provenant respectivement d'un collectif de chercheurs des universités Princeton et Stanford.

110. Il faut ajouter à cela des questions d'interprétation des modalités de la licence, car la quasi-totalité de ces licences ne contient pas de clause de loi applicable ni de forum. Il est donc possible de penser qu'en présence d'un conflit, des questions de droit international privé pourraient ajouter à la complexité de l'analyse.

111. Il faut aussi mentionner que, dans un contexte universitaire, ces licences, à part leur aspect non-commercial, sont généralement rédigées intentionnellement de façon peu restrictive, et ce, dans le but de disséminer la connaissance afin de faire avancer la science (voir la croissance du mouvement de « science ouverte »).

112. *IDCC18*, *supra*, note 102.

essentielles). De façon plus spécifique, une demande d'ajout de cette clause est souvent omise afin d'éviter de ralentir le processus de négociation avec les propriétaires de ces bases de données, ou pire, de risquer de perdre l'accès à celles-ci. De plus, la plupart de ces licences sont protégées par des mesures techniques de protection empêchant bien souvent la fouille de textes, et ce, peu importe que l'utilisation soit équitable ou non en vertu de la Loi.

Dans le deuxième cas, l'ambiguïté des droits sur les composantes de ces bases de données est un problème bien réel. Par exemple, certaines bases de données incluent un avertissement clair quant au fait qu'ils ne détiennent pas nécessairement les droits des composantes de celles-ci[113]. D'autres demeurent silencieuses sur ce point[114]. Par ailleurs, dans certains cas, l'ambiguïté quant aux usages permis des données comporte également un risque en matière de lois sur les renseignements personnels, en plus des considérations liées au droit d'auteur. Par exemple, l'incertitude quant à l'obtention de l'accord des auteurs de textes à faire partie de la base de données, ou encore, dans le cas d'une photo, l'absence de consentement à faire partie d'une compilation de données pour l'entraînement d'un algorithme à des fins particulières. Un exemple marquant : en matière de reconnaissance faciale, des enquêtes récentes révèlent qu'un nombre important de photos obtenues pour entraîner certains algorithmes à des fins de surveillance étaient utilisées sans consentement[115].

---

113. C'est le cas notamment de la base de données ImageNet : accessible sur <http://image-net.org/about-overview> et <http://image-net.org/download-faq>, ainsi que de la base de données COCO (Common Objects in Context), accessible sur <http://cocodataset.org/#termsofuse>.

114. Voir par exemple les jeux de données suivants en traitement du langage naturel : (1) superGLUE (accessible sur <https://super.gluebenchmark.com/faq>), qui préfère ne pas s'avancer quant aux droits de licences : « The primary SuperGLUE tasks are built on and derived from existing datasets. We refer users to the original licenses accompanying each dataset, but it is our understanding that these licenses allow for their use and redistribution in a research context. » ; (2) Large Movie Review Dataset (accessible sur <http://ai.stanford.edu/~amaas/data/sentiment/>), où aucune licence ne se trouve sur la page et (3) MultiNLI, (accessible sur <https://www.nyu.edu/projects/bowman/multinli/>), où le site Web réfère au texte de l'article publié, qui lui-même ne se prononce pas sur les droits conférés par le corpus utilisé dans le cadre de la recherche.

115. Une enquête du Washington Post révélait récemment que des agences gouvernementales utilisaient les bases de données de photos de permis de conduire à des fins de surveillance : Drew HARWELL, « FBI, ICE Find State Driver's License Photos Are a Gold Mine for Facial-Recognition Searches », *Washington Post*, 7 juillet 2019, accessible sur <https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/?noredirect=on&utm_term=.fa1b21d5b7c6>.

3.4 Les données transformées

Finalement, il est opportun de consacrer quelques mots à l'un des résultats obtenus par l'entraînement d'un modèle : les données transformées. On peut s'interroger sur la protection potentielle de ce type de données, particulièrement dans le cas où la génération de celles-ci est le résultat d'un processus créatif. En effet, on peut établir un parallèle entre la créativité reconnue dans l'arrêt *GSI* ci-dessus relativement aux données brutes des coupes sismiques et celles de données générées par un modèle qui contient des informations pertinentes permettant de découvrir un nouveau sens à certains textes, ou encore d'en dériver des apprentissages supplémentaires[116]. De plus, ces données transformées pourraient faire l'objet d'une protection en tant qu'adaptation, spécialement, comme dans *GSI*, dans un cas où les données entrantes étaient elles-mêmes originales. Une analyse sera requise afin de déterminer si le processus d'entraînement remplit dans les faits le seuil d'originalité prévu par la jurisprudence ou s'il s'agit simplement d'un processus mécanique.

CONCLUSION

À la lecture des enjeux décrits dans cet article, on constate que plusieurs aspects des nouveaux modèles d'affaires générés par l'application de l'intelligence artificielle font l'objet d'incertitudes en raison de l'environnement législatif actuel. Heureusement, les voix des acteurs concernés commencent à se faire entendre. Le récent rapport du Comité permanent de l'industrie, des sciences et de la technologie inclut une recommandation de « modifier la *Loi sur le droit d'auteur* afin de faciliter l'utilisation d'une œuvre ou d'un autre objet protégé à des fins d'analyse informationnelle »[117]. L'intervention des témoins ayant pris part au processus de consultation a mis l'accent sur l'importance de clarifier le statut des copies accessoires dérivées des processus technologiques : certains ont recommandé l'adoption d'une exception spécifique exemptant l'analyse informationnelle (une approche similaire à celle du Royaume-Uni[118], de l'Europe[119] et du

116. Vahe TSITOYAN, John DAGDELEN et Anubhav JAIN, « Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature », (2019) 571 *Nature* 95, p. 95-98.
117. COMITÉ PERMANENT DE L'INDUSTRIE, DES SCIENCES ET DE LA TECHNOLOGIE, *supra*, note 17, Recommandation 23.
118. *Copyright, Designs and Patents Act 1988*, 1988, c. 88, art. 29A. Il est à noter que cette exception ne s'applique qu'à la recherche non commerciale, ce qui limite grandement sa portée en pratique.
119. *Directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE*, art. 3 et 4.

Japon)[120], tandis que d'autres ont soulevé l'importance de réfléchir à une approche technologiquement neutre plutôt que de créer des dispositions pour chaque innovation[121].

Trois solutions possibles dans cette veine pourraient être envisagées. La première serait d'adapter l'article 30.71 sur les reproductions temporaires pour processus technologiques aux fins de l'analyse informationnelle. En modifiant le critère de la durée, il serait ainsi possible de permettre une reproduction temporaire, mais qui ne nécessite pas une destruction automatique, comme dans le cas de la diffusion en continu. La deuxième serait de reconnaître l'analyse informationnelle comme une fin permissible dans le cadre de l'utilisation équitable, au même titre que les autres fins permises. Cette seconde option permettrait de clarifier la légalité de l'usage de ces techniques, et ainsi favoriser le développement de nouvelles technologies, sans toutefois enfreindre les droits des créateurs des œuvres utilisées. Un tel changement pourrait stimuler la recherche et permettre de distiller la valeur intrinsèque d'un nombre important de données et potentiellement dévoiler un nouveau jour sur celles-ci. La troisième option nécessiterait un changement d'approche plus radical : éliminer la notion de fin permissible, ou à tout le moins modifier ces fins permissibles afin que celles-ci soient interprétées à titre indicatif et non exhaustif[122]. Cette approche permettrait de simplifier la Loi, de la rendre plus souple, plus neutre sur le plan technologique et, dans le contexte qui nous préoccupe, permettrait aux acteurs variés de l'IA d'agir de façon concurrentielle avec des pays ayant des dispositions similaires, notamment les États-Unis[123]. Notons que la deuxième et la troisième approche ne sont pas nécessairement mutuellement exclusives et pourraient être combinées afin de permettre à l'IA de se développer dans un environnement réglementaire moins contraignant et davantage adapté à la réalité commerciale[124].

---

120. *Copyright Law of Japan*, art. 30-4 (permettant l'utilisation et l'analyse de contenu protégé par droit d'auteur à des fins d'apprentissage automatique) ; 47-4 (permettant les copies électroniques accessoires) et 47-5 (permettant l'utilisation d'œuvres protégées pour la vérification de données, notamment la recherche de contenu dans des bases de données).
121. COMITÉ PERMANENT DE L'INDUSTRIE, DES SCIENCES ET DE LA TECHNOLOGIE, *supra*, note 17, « Analyse informationnelle », Recommandation 23 (commentaire de M[e] Mark Hayes).
122. *Ibid.*, Recommandation 18.
123. *Ibid.*, « Utilisation équitable ou exhaustive » (c'est notamment le commentaire du professeur Michael GEIST dans son « Mémoire présenté à l'INDU » daté du 14 décembre 2018) ; certaines décisions américaines ont confirmé que la fouille de textes pouvait faire l'objet de la protection de la doctrine de « fair use » : voir *HathiTrust* ; *A.V.* c. *Paradigm, LLC*, 569 F. 3d 630 (4th Cir. 2009).
124. WIPO, *supra*, note 2, relativement au fait que l'innovation en IA est générée principalement par le secteur privé.

Il faut finalement mentionner que, bien qu'il soit raisonnable d'envisager des exceptions législatives spécifiques dans un contexte où les usages sont stables, il est toutefois impossible pour le législateur d'anticiper toutes les exceptions possibles en période de profond changement technologique : à cet effet, une approche réglementaire permissive permettant d'évaluer les risques réels d'une technologie émergente semble plus sage que l'adoption d'un cadre juridique à visée trop étroite[125].

Une recommandation du Comité concernant les mesures techniques de protection[126] permet d'entrevoir un certain optimisme sur l'avenir du développement de l'intelligence artificielle au Canada : une réflexion sur la pertinence de celles-ci et l'équilibre à maintenir entre les intérêts des différents intervenants est cruciale. En effet, cette exception pourrait clore la boucle des incertitudes relatives à l'accès aux bases de données essentielles à l'entraînement des algorithmes et ainsi y mettre fin en permettant effectivement de se prévaloir d'une exception prévue par la Loi[127]. En somme, à la lumière du paysage législatif actuel, espérons que ces suggestions sauront être mises en œuvre dans un horizon prochain et apporteront la clarté requise afin de s'assurer que le Canada soit en mesure de se tailler une place de choix dans l'industrie de l'intelligence artificielle.

---

125. Pamela SAMUELSON, « Justifications for Copyright Limitations & Exceptions », *Copyright Law in an Age of Limitations and Exceptions*, Cambridge University Press, 2017.
126. COMITÉ PERMANENT DE L'INDUSTRIE, DES SCIENCES ET DE LA TECHNOLOGIE, *supra*, note 17, Recommandation 19.
127. Cela réglerait donc le problème soulevé par la Clinique d'intérêt public et de politique d'Internet du Canada dans son mémoire présenté au Comité permanent de l'industrie, des sciences et de la technologie, « L'équilibre comme guide », soumis le 14 décembre 2018, accessible sur <https://www.noscommunes.ca/Content/Committee/421/INDU/Brief/BR10269089/br-external/SamuelsonGlushkoCanadianInternetPolicyAndPublicInterestClinic-9960266-f.pdf>.