BASES

POUR UNE RECHERCHE INTELLIGENTE D'INFORMATION

N°425 • Mai 2024

SOMMAIRE

Le droit d'auteur face à l'IA générative

Philippe Masseron du gf2i (Groupement français de l'industrie de l'information) nous éclaire sur les enjeux cruciaux du droit d'auteur à l'ère de l'IA générative. Entre risques de prédation massive des contenus et opportunités d'accès et d'innovation portées par l'IA, les acteurs de la création et de l'information doivent se mobiliser pour défendre leurs intérêts et repenser leurs modèles de valorisation. Le rôle d'instances comme le gf2i sera clé pour peser dans les débats législatifs en cours et créer les nouveaux équilibres dans l'économie de la donnée.

Interview de Philippe Masseron (gf2i) sur les enjeux juridiques et économiques des métiers de l'information et de la donnée.

Anne-Marie Libmann

Philippe MASSERON

Expert en droit de la propriété littéraire et artistique et en gestion de droits

Après des études juridiques (DEA en Finances publiques et Fiscalité – Paris 2) et en Information-Communication (Institut Français de Presse – Paris 2), Philippe MASSERON a successivement occupé les postes de directeur juridique et directeur général au CFC (Centre Français d'exploitation du droit de Copie).

Il est administrateur du gf2i (Groupement Français de Industries de l'Information) délégué à la prospective. Il participe régulièrement aux missions du CSPLA.

Anne-Marie Libmann (AML) : *comment définissez-vous le problème de fond sur la question du droit d'auteur par rapport aux systèmes d'intelligence artificielle générative (SIAG) ?*

Philippe Masseron (PM) : Nous assistons à une réaccélération et une amplification massive du phénomène de scraping, fouille et crawling des données. Cela n'est pas nouveau, nous avons déjà été confrontés à des situations similaires par le passé qui ont suscité de vives inquiétudes, comme lors du lancement de Google Books ou des grands programmes de numérisation des bibliothèques. Mais l'échelle et la vitesse ont changé. Aujourd'hui, il existe d'immenses bases de contenus protégés accessibles, notamment

Le droit d'auteur face à l'IA générative *suite*

dans le domaine scientifique avec par exemple Sci-Hub qui rassemble des millions de documents en infraction avec le droit d'auteur.

Le problème posé par les IA génératives se situe à deux niveaux. En amont, il y a l'exploitation des œuvres existantes pour entraîner les modèles. En aval se pose la question de la protection par le droit d'auteur des productions de ces IA. Aux États-Unis, le Copyright Office a refusé d'attribuer un copyright à la plupart des créations d'IA qui lui ont été soumises. En Europe, il n'y a pas encore de jurisprudence établie. Mais il y a un risque évident de destruction massive de valeur et de concurrence déloyale pour les industries culturelles. Imaginez un livre généré par une IA et vendu sur Amazon qui bénéficierait de tous les avantages fiscaux et juridiques du livre sans rémunérer les auteurs dont les œuvres auraient servi à son entraînement..

AML : *Peut-on considérer que les productions issues d'IA entraînées sur des contenus non autorisés relèvent d'une forme de recel ? Cette question se pose par exemple pour les livrables de veille réalisés par les services de veille et analyse, ou encore cabinets de conseil spécialisés dans la veille et l'intelligence économique. Existe-t-il un risque juridique ?*

PM : C'est une question complexe qui mérite d'être creusée, mais je n'ai pas de réponse définitive à ce stade. La qualification de recel suppose un élément intentionnel. Faudrait-il démontrer que l'utilisateur de l'IA avait conscience que les données d'entraînement contenaient des œuvres protégées ? Il faudrait aussi être en mesure d'identifier précisément les œuvres utilisées, ce qui n'est pas évident au vu des nombreuses copies et des transformations subies.

Des techniques existent déjà pour détecter les contenus générés par IA, notamment dans le monde éducatif qui est confronté au plagiat. Mais elles ont leurs limites. Le principe de transparence figurant dans le projet d'***AIAct*** européen se veut une réponse, mais il est insuffisant, car il ne descend pas au niveau de l'identification d'œuvres individuelles au sein des jeux de données d'entraînement.

En l'état, la charge de la protection pèse donc principalement sur les ayants droit qui doivent s'équiper d'outils de détection, de marquage et de traçage. C'est coûteux et chronophage. L'*opt out* massif est une étape primordiale pour poser des barrières juridiques et ouvrir la voie à des négociations avec les exploitants d'IA. Certains grands médias comme **Le Monde** ont déjà conclu des accords avec **OpenAI**, mais la portée et le contenu de ces deals restent confidentiels.

AML : *Comment pèsera la question du référencement des sources utilisées par les IA dans les chatbots et moteurs de recherche ? Si le référencement progresse dans le bon sens, ne risque-t-on pas de réduire le problème de la juste rémunération au débat classique entre visibilité et monétisation des contenus, avec au final une perte significative dans la valorisation du travail des éditeurs de contenus ?*

PM : C'est même pire que cela, car la convergence qui s'opère entre IA et moteurs de recherche, notamment chez Google, fait peser un risque majeur sur le trafic des éditeurs de contenus. Si les liens de Google sont de plus en plus souvent servis par de l'IA au détriment des sites sources, cela pourrait avoir des conséquences dramatiques sur l'audience et donc le modèle économique des éditeurs qui en dépendent à 80 % ou plus.

D'où l'importance de l'*opt out* massif et des barrières juridiques, même si on peut penser que Google aura les moyens techniques de contourner en alimentant son IA par d'autres biais. Les éditeurs doivent aussi investir urgemment dans les outils d'analyse de logs, de marquage et de traçage des contenus, même si le coût est élevé. Cet investissement devra être pris en compte dans les futures négociations de licences.

Une action collective, portée par exemple par les organismes de gestion collective, serait sans doute profitable pour mutualiser les coûts et peser dans les rapports de force. Mais c'est compliqué dans la presse où la culture de la gestion collective est peu développée, contrairement à d'autres secteurs comme la musique. Chacun préfère souvent négocier dans son coin, les plus gros ayant les moyens de conclure des deals individuels.

AML : *L'IA n'est-elle pas le révélateur d'une crise de longue date de la propriété intellectuelle liée au Web et*

médias sociaux, avec le pillage des données initié par Google, puis amplifié par les réseaux sociaux sur les données personnelles ou encore LinkedIn sur les données professionnelles ?

PM : Il était difficile d'avoir une vision claire dès le départ. Les éditeurs se sont longtemps sentis coincés entre leur besoin de visibilité apportée par le référencement et la protection de leurs contenus. Avec l'effondrement progressif des revenus publicitaires, le rapport de force a changé, mais il était sans doute déjà trop tard.

L'IA marque une étape supplémentaire dans ce phénomène prédateur, mais on ne peut pas dire qu'elle le fait naître. C'est une forme d'accélération et de changement d'échelle, permise par la numérisation massive de ces dernières décennies et la concentration du secteur numérique autour de quelques très grands acteurs capables de se lancer dans une course à l'armement technologique.

AML : *Quels risques et opportunités voyez-vous pour les métiers de l'information (journalistes, veilleurs, documentalistes…), notamment sur la problématique du droit d'auteur lié aux résumés automatisés et autres « productions artificielles » ?*

PM : Je ne pense pas que l'IA changera fondamentalement la donne, mais il faut être vigilant. Pour les professionnels de l'information, c'est surtout une opportunité, car les outils d'IA vont permettre d'automatiser toute une partie du travail de tri, de classification et de synthèse de l'information. La valeur ajoutée se déplacera encore plus vers la pertinence de la sélection, la qualité de l'analyse et de la mise en perspective.

Il ne faut donc pas rejeter l'IA, mais réfléchir à la complémentarité intelligente avec le facteur humain. La question du résumé automatique, et de son statut par rapport au droit d'auteur de l'œuvre originale, n'est pas nouvelle. Le résumé ne permet pas, en principe, de se passer du document primaire s'il est bien fait. C'est un vieux débat, comme celui sur l'impact des panoramas de presse qui n'ont finalement pas fait disparaître les abonnements.

De même, le métier de documentaliste est toujours là malgré la disparition des centres de documentation physiques dans beaucoup d'organisations. La fonction s'est transformée et décentralisée au sein des équipes, au plus près des métiers, mais elle reste indispensable. Le vrai défi est de gérer l'infobésité croissante et d'extraire de la valeur de la masse exponentielle des données. L'IA peut aider à relever ce défi.

AML : *Pouvez-vous décrire le* rôle ***et actions du gf2i pour défendre*** *les intérêts des acteurs de l'industrie de l'information dans ce contexte de bouleversement* ?

PM : Le gf2i a vocation à sensibiliser ses membres aux enjeux et à promouvoir l'émergence de marchés pour les différents types de données, au-delà des seuls contenus éditoriaux.

L'enjeu majeur pour nous est de structurer de véritables marchés pour les différents types de données. Des embryons existent déjà, mais tous les acteurs n'en ont pas encore pleinement conscience. Le mouvement de l'*open data* a un peu brouillé les pistes, avec des effets positifs en termes d'accès, mais aussi un appauvrissement côté public, avec une perte de qualité par manque de moyens pour maintenir et mettre à jour les jeux de données ouverts.

AML : *Comment structurer de véritables marchés de données ?*

PM : L'enjeu porte sur tous les domaines : données juridiques, géographiques, de santé, financières, etc. C'est un immense défi de structuration qui nécessite de trouver les bons modèles économiques et les bonnes formules de licences adaptées aux différents maillons de la chaîne de valeur. Le droit d'auteur a toute sa place, mais il faut aussi valoriser les investissements dans la qualité, la mise à jour, l'enrichissement, le croisement des données.

En ce sens, un service de veille et de documentation interne peut aussi se penser comme un producteur de données à valoriser, pas seulement comme un consommateur. C'est un changement de perspective à opérer.

En conclusion, on voit que malgré les bouleversements technologiques, les questions de fond sur la protection et la valorisation des données et des contenus restent assez similaires. Le gf2i est mobilisé de longue date sur ces enjeux et entend bien continuer à peser dans les débats actuels, comme dans le cadre des missions en cours du CSPLA (Conseil Supérieur de la Propriété Littéraire et Artistique) et de la mission parlementaire sur ces sujets. Il en va de la survie de pans entiers de l'économie de l'immatériel.