[Dossier IA générative] - Quelles régulations pour la conception des IA génératives ?

Rédigé par Erevan Malroux - 26 avril 2023

La popularité spectaculaire de l'interface ChatGPT a suscité une série de débats juridiques, au point d'appeler à l'adoption d'une réglementation dédiée.

Pour autant, différentes réglementations existent et s'appliquent déjà à de tels systèmes, même si des clarifications sont nécessaires.

Cet article a vocation à explorer certaines règles structurantes (comme le droit d'auteur ou la protection des données personnelles) qui s'imposent à la conception et à l'utilisation de systèmes d'IA générant du contenu tel que du texte ou des images, indépendamment du projet de règlement européen sur l'IA actuellement en cours d'élaboration.

Sommaire :

« *Non, ce n'est pas trop tôt.* » Voilà ce qu'a répondu Mira Murati, directrice technique d'OpenAI, **à un journaliste du Time qui lui demandait si l'adoption d'une réglementation étatique** ne risquerait pas de ralentir l'innovation en matière d'IA générative comme ChatGPT (traduction LINC). Cette réponse fait notamment écho à une série d'initiatives, il y a déjà plusieurs années, d'un **certain nombre d'acteurs qui se dotaient de chartes éthiques** pour le développement de l'IA.

Comme souvent lorsqu'une industrie commence à susciter des craintes, ses acteurs y répondent dans un premier temps en proposant une forme d'autorégulation. C'était notamment le cas d'OpenAI, initialement créée comme une association à but non lucratif et open source, qui a consenti d'importants efforts de modération s'agissant de la mise à disposition de son système ChatGPT (par exemple, en lui apprenant à refuser de répondre à certaines questions, bien que ce filtre soit imparfait).

Plus récemment, une **initiative collective d'autorégulation a vu le jour**, potentiellement plus contraignante pour les acteurs. Mais cela n'exclut pas un autre risque non négligeable, celui que les standards soient définis par un trop petit nombre d'acteurs privés. Indépendamment des efforts consentis, définir des critères de modération n'est jamais neutre, et révèle nécessairement des conceptions du monde et de différentes valeurs. Pour reprendre l'exemple de ChatGPT, utilisé par plus de cent millions d'utilisateurs dans le monde, c'est une société américaine de quelques centaines de salariés qui a défini les catégories de questions et de réponses non autorisées (OpenAI a d'ailleurs **révélé avoir consenti encore plus d'efforts dans la modération** de son nouveau modèle GPT4 qu'elle ne l'avait fait pour le modèle GPT3.5 sur lequel reposait ChatGPT initialement).

Pour ces raisons, de nombreux spécialistes, y compris au sein d'OpenAI, appellent à l'adoption de réglementations dédiées à ces systèmes d'IA, pour que les règles ne soient pas seulement définies par les entreprises. Encore plus récemment, **un appel à moratoire a été lancé en la matière**.

Pour autant, si ces débats sont nécessaires, ils ne doivent pas faire oublier les règles existantes qui s'appliquent déjà à ces dispositifs. Cet article aborde le droit d'auteur et le droit de la protection des données personnelles qui sont des réglementations structurantes pour la conception de ces systèmes. Il n'a cependant pas vocation à aborder la question de l'imputabilité des manquements éventuels à ces règles, ni à détailler les enjeux liés à la génération de contenus interdits par la loi comme en matière de désinformation (ces usages étant davantage abordés **dans l'article 3 du dossier**).

Pourquoi le vide juridique est un mythe – l'exemple du droit d'auteur

Juridiquement, le vide juridique n'existe pas

Parler de vide juridique est en réalité un abus de langage. En effet, même dans les cas où il n'existe pas de loi ou de règlement spécifique qui s'applique à une situation donnée, des principes généraux permettent toujours de l'appréhender juridiquement, que ce soit pour interdire une pratique, ou pour encadrer son usage. Lorsqu'une pratique n'est contraire ni à aucune règle spécifique, ni à un principe juridique général, la liberté prévaut et elle est permise. Un juge saisi d'une affaire a d'ailleurs l'obligation de se prononcer, que le droit soit adapté ou non, au risque de commettre **un déni de justice**. Cela est aussi valable pour les systèmes d'IA, en particulier en France et en Europe, où il existe déjà plusieurs réglementations directement en lien avec le développement et l'utilisation de tels dispositifs.

En réalité, l'enjeu n'est pas l'absence de réponse en droit, mais plutôt le doute qui peut exister sur le sens de cette réponse (il serait alors plus juste de parler de flou juridique, notamment en l'absence de précédents), ou le fait que la réponse n'est pas jugée satisfaisante (par exemple lorsqu'une pratique est permise, alors que la société estime qu'elle devrait être plus strictement encadrée). Dans ce dernier cas, il convient alors d'envisager de changer le droit, par exemple en adoptant une réglementation spécifique.

La question du droit d'auteur est sans doute l'un des exemples les plus parlants sur l'absence de vide juridique.

Conception et usage de modèles d'IA, les contours du plagiat

Sans aller jusqu'à parler de flou artistique, de nombreux commentateurs se sont interrogés sur le respect du droit de la propriété intellectuelle par les créateurs de systèmes d'IA générative depuis leur adoption massive au cours des derniers mois, en particulier s'agissant des systèmes permettant de générer des images.

Comme la plupart des IA génératives, ce modèle repose sur l'application de techniques de deep-learning à des quantités colossales de données, le plus souvent collectées sur internet. De premières questions juridiques se posent alors : le modèle de langage n'aurait-il pas violé le droit d'auteur en se construisant sur la base de contenus protégés ? N'y a-t-il pas un risque, en utilisant de tels systèmes, de se rendre coupable de plagiat si le résultat « s'inspire » trop directement d'une œuvre utilisée pour entraîner le modèle ?

Ces questions posent de grands enjeux. Philosophiquement, cela revient presque à interroger ce qu'est une œuvre (« peut-on parler d'œuvre sans ouvrage » serait d'ailleurs digne d'un sujet de baccalauréat). Economiquement, l'enjeu est très concret puisqu'il pose la question de l'éventuelle rémunération du travail publié sur internet, alors que ces dispositifs qui les réutilisent font **l'objet d'une industrialisation croissante**. Un autre enjeu de société est aussi d'interroger la pertinence des règles en la matière, puisque les juges devront trancher la question en appliquant les règles de droit actuellement en vigueur.

Si aucun juge ne s'est encore prononcé sur le sujet, **plusieurs actions en justice** ont déjà été intentées à l'encontre de concepteurs de systèmes d'IA dans plusieurs juridictions (il s'agit en particulier d'IA génératives d'images, accusées de violer le copyright américain ou anglais). Cela concerne également les systèmes reposant sur des modèles de langage (voir article 1), puisque les textes peuvent aussi être considérés comme des œuvres protégés par le droit d'auteur (**la notion d'œuvre étant potentiellement très large**, au-delà des poèmes, romans et autres œuvres littéraires, elle comprendrait aussi des billets postés sur un réseau social, des articles de blog, des sites web, ou encore des codes informatiques postés sur un forum).

S'agissant de l'utilisation de contenu pour créer un modèle d'IA, le droit européen fournit une réponse relativement précise. En effet, une directive de 2019 **a introduit dans le droit de l'Union européenne une exception** pour la fouille de texte et de données (qui est l'une des techniques utilisées pour concevoir de tels systèmes). **Cette exception** permet de collecter des contenus publiquement accessibles sans demander le consentement de leurs auteurs, y compris pour une exploitation commerciale, à condition que les auteurs en question aient pu s'opposer à cette réutilisation de leurs œuvres. Cette faculté d'opposition, encore assez méconnue par les auteurs, pourra poser des questions quant à ses modalités d'application (comment est-il possible de s'opposer, comment les concepteurs de systèmes d'IA doivent prendre en compte ces oppositions, sera-t-il aussi facile de faire valoir son opposition pour un auteur indépendant ou des organismes plus structurés ?).

Malgré ces interrogations pratiques, il semble que le droit européen avait déjà anticipé ce cas, en particulier comparé au droit américain, dont l'application de **l'exception dite de « fair use » au copyright américain** fait davantage débat s'agissant de l'entraînement des modèles à partir d'œuvres protégées.

Reste alors la question du respect du droit d'auteur, non par l'entraînement du modèle mais par la génération d'images ou de textes par de tels système. En effet, une contrefaçon pourrait être caractérisée si le résultat était trop directement « inspiré » d'une œuvre protégée. Pour déterminer s'il s'agit d'une violation du droit d'auteur, une analyse au cas par cas sera nécessaire.

Là encore, les règles qui s'appliqueront à un éventuel litige sont déjà déterminées, bien qu'elles ne soient pas les mêmes partout dans le monde.

En France, il faut vérifier que l'œuvre d'origine est une création protégée (c'est-à-dire s'il s'agit d'une œuvre originale, peu importe son genre, sa forme d'expression, son mérite ou sa destination) et qu'elle a été reproduite sans l'autorisation de son auteur (y compris de manière partielle ou par l'emprunt de ses caractéristiques originales). Le propre de ces systèmes d'IA est d'apprendre sur la base de quantités phénoménales de contenus, ce qui devrait - au moins en théorie - réduire le risque que ses résultats soient assimilables à du plagiat. Pour autant, ce risque n'est pas négligeable, en particulier en fonction des invites de commande (ou prompts) de l'utilisateur (se posera alors une question de la responsabilité) et des éléments originaux susceptibles d'avoir été appris par le modèle.

Il est à noter que le droit américain prend également en compte l'originalité et la créativité d'un contenu pour pouvoir le protéger par le copyright. **L'office américain du copyright s'est d'ailleurs récemment prononcé** sur la réalisation d'une bande dessinée à l'aide d'une IA générative d'images (jugeant le processus insuffisamment créatif en l'espèce, du moins s'agissant des illustrations générées par le système, en raison de son caractère aléatoire).

Si d'autres réglementations sont potentiellement en tension avec la conception ou l'utilisation de tels systèmes d'IA (ex : le respect des conditions générales des sites web, **la protection du secret des affaires**, auxquels pourrait s'ajouter la sanction du parasitisme), la question du droit d'auteur est l'une des plus médiatiques et potentiellement l'une des plus structurantes.

Il en va de même pour la protection des données personnelles qui trouve elle aussi pleinement à s'appliquer.

La protection des données traitées par ChatGPT

Appliqués concrètement au cas de ChatGPT, les enjeux de protection des données personnelles se placent essentiellement à deux niveaux :

- Celui de la conception du modèle de langage sous-jacent qui utilise structurellement de très nombreuses données collectées sur internet ;

- Celui de l'utilisation du système ChatGPT qui traite les données d'utilisateurs et peut les réutiliser pour améliorer le service ou en développer de nouveaux.

Le RGPD prévoit un certain nombre de principes et d'obligations pertinents qui s'appliquent à ces différentes phases.

« IA ouvre-toi » ou le sésame du droit d'accès ?

Comme rappelé plus haut, les modèles d'IA générative reposent sur la collecte et l'utilisation de quantités colossales de données (dont certaines sont personnelles), le plus souvent sur internet. Si certaines de vos données sont en ligne (par exemple sur des réseaux comme Twitter ou Reddit), il est possible qu'elles aient servi à entraîner le modèle de langage sur lequel repose ChatGPT.

Comme aux prémices des moteurs de recherche sur le web, un premier réflexe serait de demander directement à ChatGPT s'il vous « connaît » dans son interface.

Mais comme il s'agit d'un modèle statistique et non d'un moteur de recherche sa réponse, même positive, ne serait pas très utile. En réalité, cela révèlerait surtout sa capacité à prédire une suite cohérente de lettres, mots ou bouts de phrases associés à la suite de lettre qui compose votre nom. Pire, **s'agissant d'un modèle probabiliste**, sa réponse pourrait ne pas être toujours la même, voire être fausse.

Vous n'apprendrez donc pas nécessairement ce qu'il « sait » ou ce qu'il a pu « lire » sur vous, et encore moins sur quels sites web, puisque **ChatGPT ne cite pas ses sources dans ses réponses**.

Rappelons alors que le RGPD s'applique à la collecte et à la réutilisation des données personnelles, quand bien même ces données ont été publiquement accessibles sur internet. Il prévoit notamment que ces données doivent être traitées de manière licite, loyale et transparente.

Ce principe de transparence est fondamental dans la protection des données personnelles, en particulier dans la régulation des systèmes d'IA. Il impose au responsable de la collecte des données sur internet, de leur éventuelle annotation, et de leur utilisation à des fins d'entraînement d'un modèle d'IA, de fournir une information pertinente sur les conditions dans lesquelles ces données sont traitées, et notamment sur les sources de la collecte. Dans la même logique, le RGPD prévoit un droit d'accès, qui permet à tout individu de savoir si un organisme traite des données le concernant, d'en obtenir une copie le cas échéant, et de recevoir des informations plus précises sur la manière dont elles sont traitées.

Le droit d'accès est une composante essentielle de la transparence, et permet notamment d'orienter ou d'exercer d'autres droits prévus par le RGPD, tant à l'égard du traitement des données qui auraient pu être collectées en ligne pour concevoir le modèle d'IA, qu'à l'égard du traitement des données fournies par chaque utilisateur qui utilise son interface conversationnelle.

« AI-je mon mot à dire sur l'utilisation de mes données ? » ou l'exercice des autres droits prévus par le RGPD

Le RGPD n'impose pas toujours le consentement des personnes pour traiter leurs données. D'autres garanties existent et permettent d'assurer la protection des données personnelles sans demander l'accord de chaque individu. Ces alternatives valent aussi dans le cadre de la constitution et de l'exploitation de systèmes d'IA générative.

La philosophie du RGPD est de permettre aux personnes de conserver la maîtrise de leurs données à travers une série de droits qui leur sont accordés. C'est notamment le cas du droit d'opposition, qui permet de s'opposer au traitement de ses données par un organisme pour un motif légitime et qui peut s'exercer à tout moment, mais aussi du droit à l'effacement ou du droit à la rectification des données.

Lors de la phase de conception du modèle de langage, la collecte et l'utilisation de données publiquement accessibles posent de nombreuses questions quant à leur licéité (les mesures et garanties en place sont-elles suffisantes pour se passer du consentement des personnes concernées ?), ou la possibilité pour chaque individu d'exercer ses droits sur les données personnelles qui le concernent, et qui seraient incluses dans la base d'entraînement.

S'agissant de la phase d'utilisation de ChatGPT, l'interface conversationnelle peut amener les utilisateurs et le grand public à confier des données **potentiellement sensibles ou confidentielles** au système (bien qu'un avertissement de l'interface elle-même appelle à ne pas fournir de telles données). A cet égard, il a lieu de noter que les conditions générales d'utilisation de la version gratuite de ChatGPT prévoient qu'OpenAI puisse réutiliser les données des utilisateurs pour améliorer le système, ou en développer de nouveaux (**un mécanisme d'opposition à cette réutilisation semble prévu en l'espèce**).

Là encore, le RGPD encadre les conditions de réutilisation des données fournies par les utilisateurs, à travers une série de droits accordés aux personnes, mais aussi à travers d'autres obligations que le responsable du traitement doit anticiper.

L'AIPD, une analyse qui porte bien son nom, ou le principe d'accountability

Le RGPD impose à tout responsable du traitement de respecter ses obligations, mais aussi de pouvoir démontrer sa conformité (il s'agit du principe de responsabilité, aussi appelé « accountability »). Chaque organisme traitant des données personnelles doit mettre en œuvre des mesures plus ou moins importantes en fonction des risques encourus par les personnes concernées, et ce par défaut et dès la conception des traitements.

A cet égard, **l'analyse d'impact sur la protection des données** (ou AIPD) est un outil prévu par le RGPD, qui aide à penser, mais aussi à démontrer la conformité d'un traitement. L'AIPD est d'ailleurs obligatoire pour les traitements de données personnelles susceptibles d'engendrer un risque élevé pour les droits et libertés des personnes concernées (par exemple en cas d'usage innovant sur des données collectées à large échelle). Il est à noter que **le droit à la protection de ses données personnelles est distinct du droit au respect de sa vie privée**, puisqu'il a vocation à protéger d'autres droits et libertés.

L'AIPD doit donc prendre en compte les risques d'atteinte à la vie privée, mais aussi des risques à d'autres droits et libertés. Il est possible de citer les enjeux de qualité, de quantité et de pertinence des données (dont le risque de collecter indûment des données sensibles) ou encore les risques en matière de sécurité des données (qu'il s'agisse **des données contenues dans le modèle** ou des données transmises par les utilisateurs de l'interface), mais aussi **les risques de biais que pourraient développer le modèle**, notamment en raison des biais déjà présents dans les données servant à son entraînement, ou des biais dans le choix des données sélectionnées.

La philosophie de l'AIPD et son approche par les risques, sont d'ailleurs au cœur du projet de règlement européen sur l'IA (RIA) actuellement en cours d'élaboration.

La décision de l'autorité italienne du 31 mars 2023

Le **31 mars 2023**, l'homologue italienne de la CNIL a adopté une décision d'urgence (**voir ici**) à l'encontre de la société américaine OpenAI L.L.C. qui développe et gère le service ChatGPT. Cette décision interdit temporairement à OpenAI de traiter des données des utilisateurs italiens.

L'autorité italienne a considéré qu'il existait un doute suffisant sur la conformité au RGPD de ChatGPT pour prononcer cette mesure d'urgence, notamment au regard de l'information des utilisateurs dont les données ont servi à l'apprentissage du modèle, à la base légale du traitement de constitution du modèle, au caractère inexact de certaines données et à l'absence de tout mécanisme de vérification de l'âge sur un service présenté comme étant réservé aux personnes âgées de plus de 13 ans.

Le **12 avril 2023**, après une série de discussions avec les représentants d'OpenAI, l'autorité italienne a indiqué certaines exigences à remplir d'ici le 30 avril 2023 (en matière de transparence, de droits des parties intéressées et de base juridique du traitement effectué par ChatGPT - **à lire ici**) pour suspendre son interdiction temporaire.

Le **13 avril 2023**, le Comité européen de protection des données, qui réunit la CNIL et l'ensemble de ses homologues européens, a décidé de lancer un groupe de travail sur ChatGPT (**à lire ici**). Son objectif est de favoriser la coopération et l'échange d'informations sur d'éventuelles initiatives visant à assurer l'application du RGPD par les différentes autorités de régulation.

Le **28 avril 2023**, l'autorité italienne recense neuf mesures prises par OpenAI permettant ainsi le rétablissement de l'accès à ChatGPT pour les internautes basés en Italie (**voir ici**). Ces mesures, qui s'appliquent dans toute l'Union Européenne, comprennent en particulier : une **information** détaillant les traitements mis en œuvre, la possibilité de **s'opposer au traitement de ses données** pour les utilisateurs et non-utilisateurs (via **un formulaire** ou par **email**), l'introduction de **mesures d'effacement des données inexactes** (sachant que la correction des données apparaît impossible aujourd'hui), la clarification de la façon dont les **données des utilisateurs peuvent être réutilisées** à des fins d'amélioration de l'algorithme sans préjudice de pouvoir s'y opposer, la mise en œuvre d'un **mécanisme de déclaration de l'âge** (avec interdiction pour les utilisateurs de moins de 18 ans sauf pour les mineurs agées entre 13 et 18 ans bénéficiant du consentement de leurs parents).

Vers une nouvelle approche de la régulation des systèmes d'IA ?

Une nouvelle manière d'appréhender la protection des données personnelles ?

Comme nous venons de le voir, la réglementation en matière de protection des données repose sur des grands principes très robustes. Le RGPD reste donc non seulement applicable, mais tout à fait pertinent, pour les traitements des systèmes d'IA générative qui requièrent de grandes quantités de données, souvent personnelles. **Une récente étude du Conseil d'État** souligne d'ailleurs « la très forte adhérence entre la régulation des systèmes d'IA [à venir] et celle des données, en particulier des données à caractère personnel ».

Toutefois, si la collecte de données en ligne ou la réutilisation des données des utilisateurs pour améliorer ou développer de nouveaux services ne sont pas inédits, la conception de modèles d'IA génératives d'une telle ampleur pose des enjeux juridiques et techniques nouveaux. Par exemple, s'agissant des modèles statistiques au cœur des IA génératives, garantir le respect des droits des personnes pose des questions nouvelles, telles que **la possibilité de retrouver des données du jeu d'entrainement**, ou **d'exercer ses droits directement sur le modèle**. La CNIL se penche sur ces questions depuis plus d'un an et a déjà eu l'occasion de publier **des ressources sur le sujet**, en particulier à destination des professionnels.

Par ailleurs, face aux avancées technologiques spectaculaires de l'IA et pour répondre à ces enjeux de société, **la CNIL a récemment créé un service de l'intelligence artificielle**. Dans ce contexte, elle devra analyser comment le RGPD peut venir encadrer le développement d'IA génératives et leurs usages afin d'en assurer une régulation équilibrée et permettre aux organismes, aux personnes concernées, et à la CNIL de maîtriser les risques pour la vie privée. Des travaux en ce sens seront conduits au cours de l'année 2023 pour y apporter de premières réponses.

Le projet de règlement sur l'IA, en complément au RGPD

Les travaux autour du projet de règlement sur l'IA pourraient laisser croire que les réglementations actuelles sont obsolètes ou dépassées. Pourtant, ce projet de règlement n'a pas vocation à remplacer le RGPD pour ces systèmes mais bien à le compléter.

A ce stade, le projet de RIA précise d'ailleurs un certain nombre de principes déjà présents dans le RGPD, de manière plus générale. C'est notamment le cas du principe de transparence et de la prise en compte des risques en amont de la conception des systèmes. De plus, le règlement IA a vocation à s'appliquer d'abord aux « fournisseurs de solution » d'IA souhaitant accéder au marché européen, là où le RGPD concerne principalement des utilisateurs de solution qui sont responsables des traitements qu'ils achètent ou mettent en œuvre.

Il se montre en revanche plus précis dans son approche par les risques notamment en proposant différentes catégories de risque pour chaque usage, les mesures à prendre étant alors plus ou moins contraignantes. Le projet de règlement distingue notamment les systèmes d'IA présentant des risques inacceptables (lesquels sont interdits), les systèmes à « haut risque » (soumis à des exigences élevées) et les systèmes à risque « limité » (soumis à des exigences minimales, notamment de transparence).

A cet égard, la question se pose de trouver le bon niveau d'encadrement s'agissant des systèmes **d'IA dits à usage général**, dont les IA génératives sont une sous-catégorie. **Faut-il créer un régime sur mesure** comme cela a un temps été envisagé (notamment en raison de la difficulté d'envisager tous les usages à risques dès le stade de la conception de tels systèmes) ? **Faut-il les considérer comme des systèmes d'IA à « haut risque »**, ce qui exigerait des mesures importantes pour pallier les différents risques encourus, telles qu'une transparence accrue à l'égard des utilisateurs, une attention particulière quant aux données

alimentant le système, et un niveau élevé de robustesse, de sécurité et d'exactitude ? Au-delà du règlement IA, en France, le rapport du Comité Pilote d'Ethique du Numérique **sur les agents conversationnels** a émis 13 préconisations, 10 principes de conception et 11 questions de recherche à suivre pour ce type de systèmes. Parmi ceux-ci, « Affirmer le statut des agents conversationnels », « Réduire la projection de qualités morales sur un agent conversationnel » ou « Encadrer techniquement les deadbots » sont des préconisations dont la mise en œuvre permettrait sans doute de répondre aux enjeux précédemment cités.

Les co-législateurs européens devront répondre à ces interrogations. Certains commentateurs se demandent si les nouvelles obligations à la charge des concepteurs de tels dispositifs seront suffisantes (notamment s'agissant des obligations de transparence sur les systèmes d'IA à usage général), d'autres craignent que des obligations supplémentaires trop lourdes ne freinent l'innovation des entreprises, au profit d'acteurs étrangers.

De la même manière que le RGPD ne s'applique pas uniquement aux organismes européens, le projet de RIA s'appliquerait également aux fournisseurs étrangers qui commercialiseraient ou mettraient en service leurs systèmes d'IA sur le marché européen.

Illustration - Guilhem Vellut (Flickr)

*Article rédigé par **Erevan Malroux**, Juriste au service des affaires économiques*

VOIR PLUS D'ARTICLES DE L'AUTEUR