Protection des données et intelligence artificielle : une gouvernance indispensable

Philippe Gilliéron, le 23 octobre 2023

Cette contribution s'efforce de tirer un parallèle entre les grands principes régissant la protection des données personnelles et les exigences prévues par la Proposition UE de Règlement sur l'IA, en particulier en ce qui a trait à la gouvernance devant entourer le développement et la mise en œuvre de tels systèmes.

La protection des données est l'un des nombreux enjeux que pose le développement des systèmes d'intelligence artificielle. Compte tenu du rôle majeur que joue aujourd'hui le cadre règlementaire applicable aux données en matière d'innovation et, corollairement, sur le plan géopolitique, il n'est pas surprenant que ce cadre soit en pleine évolution, notamment au sein de l'Union européenne.

Sans prétendre ici passer en revue ce cadre en devenir, cette contribution s'interroge sur la mise en œuvre des grands principes applicables en protection des données à ces systèmes. On constatera que ces principes font à bien des égards partie intégrante des principes éthiques régissant le développement des systèmes d'intelligence artificielle que sont les notions de « *fairness* », « *accountability* » et « *transparency* » (I). Ceci fait, nous examinerons la manière dont la Proposition de Règlement sur l'IA (« Proposition »), comparée à l'approche qui prévaut pour le moment aux États-Unis, propose de gouverner le développement et la mise en œuvre de ces systèmes, notamment en ce qui a trait à la gestion des risques (II). Nous formulerons enfin quelques remarques conclusives s'agissant de l'adoption des standards et des questions de gouvernance qui entourent le déploiement de ces systèmes (III).

Le cadre de cette étude étant limité en raison de sa brièveté, le lecteur voudra bien excuser les raccourcis parfois utilisés, potentiellement source d'imprécision, le but étant avant tout de souligner les enjeux sans forcément prétendre les résoudre ici.

I. Enjeux en matière de protection des données

A. Répartition des rôles

Une première question consiste à savoir qui, dans un cas donné, sera considéré comme le

responsable de traitement, puisque c'est avant tout sur ce dernier que repose le devoir de conformité.

Au stade du développement du système, ce responsable sera à l'évidence le développeur, le client utilisateur ne jouant alors aucun rôle. C'est alors au développeur qu'appartient le devoir de s'assurer que les grands principes mentionnés ci-dessous sont respectés.

Au stade de l'utilisation du système, le responsable sera en revanche en principe l'utilisateur, le développeur n'apparaissant alors plus que comme le sous-traitant qui fournit le système à son client à l'image de n'importe quel autre prestataire. La situation peut toutefois nécessiter une appréciation différente à partir du moment où le développeur se réserve le droit d'utiliser les données ingérées par son client pour permettre l'amélioration de son algorithme. En cette hypothèse, le développeur apparaîtra comme un responsable séparé (éventuellement conjoint si le client est contractuellement en droit de profiter de ces améliorations subséquentes) ; chacun sera alors tenu d'assurer le respect des exigences posées en matière de protection des données.

Pour des raisons de place, on ne mentionnera pas ici le principe de licéité, dès lors qu'en pratique, le développeur, respectivement l'utilisateur, s'efforceront de faire valoir qu'ils détiennent un intérêt légitime sous l'angle du RGPD, respectivement un intérêt privé prépondérant à traiter de telles données sous l'angle de la LPD pour permettre la mise en œuvre du système. Si l'on peut à dire vrai s'interroger sur la pertinence systématique de ce motif justificatif, véritable fourre-tout dont le bien-fondé est de plus en plus tempéré par la CJUE, nous n'entrerons pas ici dans les détails d'une telle analyse qui mériterait une contribution en soi.

B. Principe de protection des données dès la conception

Conformément au principe de protection des données dès la conception désormais ancré aux art. 25 RGPD et 7 LPD, le plus simple consiste à ne pas ingérer dans le système de données personnelles pour l'entraîner.

De nombreuses techniques recensées par exemple par l'OCDE sont aujourd'hui envisageables, chacune présentant des avantages et des inconvénients, que l'on pense à la simple anonymisation de données, le recours à des données synthétiques, à la confidentialité différentielle (*differential privacy*), aux divers systèmes de cryptographie ou encore à des techniques comme le « *zero knowledge proofs* », à peu près intraduisible.

Outre le fait que ces technologies nécessitent une bonne gouvernance pour être mises en

œuvre dès le début du projet, leur recours peut s'avérer onéreux (par exemple en raison du volume de données à nettoyer) et, le cas échéant, ne pas permettre d'atteindre l'objectif visé (un des problèmes qui semble être évoqué s'agissant du recours à des données synthétiques, sans mentionner le fait que l'anonymisation peut s'avérer toujours plus difficile à atteindre en certains secteurs, notamment en ce qui a trait aux données de santé).

Lorsqu'une telle anonymisation s'avère impossible ou n'a pas été prise en compte dès le début comme l'exigerait pourtant le principe « *privacy by design* », se pose la question du respect des principes de finalité, de minimisation et de conservation des données.

C. Principes de finalité et de minimisation

Le principe de finalité commande que seules les données nécessaires pour atteindre le but visé soient traitées. Ce principe est ainsi étroitement lié à celui de minimisation des données.

De prime abord, le respect de ces principes apparaît raisonnablement envisageable pour les systèmes ayant une finalité clairement prédéfinie lors de leur conception. Leur respect est toutefois à double tranchant : un bon nettoyage des données pourra de prime abord permettre d'avoir des données de qualité et éviter une redondance inutile (susceptible de faciliter des attaques) ou la favorisation de biais. D'un autre côté, il peut toutefois s'avérer difficile pour le développeur d'apprécier d'emblée quelles seront le type de données nécessaires pour entraîner au mieux l'algorithme et éviter des biais, un nettoyage trop important pouvant avoir un effet contre-productif en la matière.

Assurer le respect de ces principes, même en partie, pour des modèles de fondation (« *foundational models* »), en vogue aujourd'hui et qui ont pour objectif de permettre un éventail d'utilisations aussi large que possible, apparaît encore beaucoup plus hasardeux. Est-ce à dire qu'au vu des finalités particulièrement larges envisageables de ces modèles, aucun principe de minimisation ne serait applicable puisque, de prime abord, toutes les données auraient leur rôle à jour ? Tel n'est à mon avis pas le cas, bien au contraire.

De prime abord, pour de tels systèmes (à tout le moins leur large majorité à ce jour), aucune donnée personnelle ne s'avère nécessaire pour atteindre l'objectif visé. Pour les modèles de fondation, le mot d'ordre devrait ainsi être de recourir à des techniques d'anonymisation des données ou l'interdiction de recourir à des données personnelles pour éviter une violation alors particulièrement aisée du principe de minimisation. Est-il besoin de dire qu'à l'heure où les investissements dans les modèles de fondation sont colossaux, le respect de telles exigences apparaît utopique ? Faudrait-il considérer que celui qui divulgue de manière

publique sur le Net certaines données personnelles consent implicitement à leur exploitation à des fins d'entraînement des algorithmes ? Ce débat n'est pas sans rappeler celui qui avait animé les spécialistes du droit d'auteur lors de l'avènement du *World Wide Web* à la fin des années 90 et la question de savoir si la mise en ligne de contenus protégés par les titulaires de droits impliquait une licence autorisant les tiers à lier ces contenus ; au final, l'introduction de dispositions légales autorisant les copies provisoires rendues techniquement nécessaires avait mis un terme au débat. En ira-t-il de même de l'entraînement des algorithmes des modèles de fondation au moyen de données rendues publiques ? Le débat est ouvert.

À ce jour, il n'est en tous les cas pas certain que la simple possibilité conférée aux utilisateurs recourant à de tels systèmes de faire du « *opt out* » pour éviter que leurs données ne soient ingérées dans le système, comme le permettent de nombreux systèmes en leur version payante, aussi dite « *business* » ou « *enterprise* », suffisent à satisfaire au principe de minimisation (étant précisé que ces utilisateurs devraient éviter que ces données ne soient personnelles, ce que les conditions générales prévoient du reste en principe).

Là encore, l'établissement d'une gouvernance adéquate apparaît dès lors comme primordial pour se poser les bonnes questions dès le départ.

D. Durée de conservation

En principe, les données ne doivent pas être conservées plus longtemps que nécessaire. La question du « nécessaire » est bien souvent largement interprétée comme permettant une conservation des données aussi longtemps que la finalité n'est pas atteinte. Or, cette finalité peut être définie assez largement, telle l'amélioration de l'algorithme, vidant au final cette exigence de son sens.

On peut dès lors douter qu'une telle durée soit acceptable s'agissant du traitement de données personnelles à des fins d'améliorer *ad aeternam* un algorithme, ce d'autant si leur l'utilité est d'emblée discutable.

À partir du moment où il peut s'avérer techniquement particulièrement compliqué de ne nettoyer que les données personnelles des données d'entraînement une fois ingérées dans le système, on y voit là une autre raison pour éviter d'emblée le recours à de telles données, sauf si la finalité du système en exige le recours.

Si le respect des points susmentionnés est important, l'un des points les plus cardinaux en lien avec ces systèmes réside dans la mise sur pied de mesures techniques et organisation-

nelles adéquates pour assurer la confidentialité, l'intégrité et la disponibilité des données exploitées, tel que prévu aux art. 32 RGPD et 8 LPD.

II. Mesures techniques et organisationnelles : l'appréciation du risque

Les malveillances et inconvénients susceptibles d'affecter de tels systèmes sont connus (du moins en partie), qu'ils aient trait au manque de qualité des données ingérées par le système (redondance, biais) aboutissant à des résultats inappropriés, ou à une exploitation indue de ces données dues à des manquements sur le plan sécuritaire (qu'il s'agisse de « *privacy attacks* », « *adversarial attacks* » ou encore « *poisoning attacks* » pour prendre celles répertoriées par le *Bundesamt für Sicherheit in der Informationstechnik* en matière de *Large Language Models*).

À ce titre, l'exécution d'une analyse d'impact quant à l'éventuelle survenance des risques liés à la mise sur pied d'un système donné risque fort de devenir un standard sur le plan international comme en témoignent les approches européenne et américaine résumées ci-dessous :

A. La Proposition de Règlement sur l'intelligence artificielle

Sur le plan européen tout d'abord, la Proposition prévoit un cadre de gouvernance rigoureux pour les systèmes dits à « risques élevés », qui ne s'appliquent certes pas dans leur intégralité aux autres systèmes, lesquels n'en seront pas moins soumis dans leur développement et leur exploitation à certaines exigences, notamment de transparence s'ils présentent un risque limité.

Est ainsi prévu en premier lieu pour les systèmes à « risques élevés » à l'art. 9 de la Proposition la mise sur pied d'un système de gestion des risques visant à identifier les risques connus ou prévisibles ou susceptibles d'apparaître tant dans le cadre d'une bonne ou mauvaise utilisation ainsi que les mesures envisagées pour réduire ou atténuer à tout le moins le risque résiduel.

Cette méthodologie n'est pas sans rappeler celle de l'analyse d'impact en matière de protection des données prévue à l'art. 35 RGPD. Il n'est donc guère surprenant que de nombreuses entreprises envisagent d'intégrer une telle analyse à celle de l'art. 35 RGPD, 40% des entreprises consultées lors d'une étude menée par l'IAPP ayant de surcroît prévu d'intégrer de telles analyses algorithmiques à leur analyse d'impact en matière de protection des données. Parmi les risques devant être pris en compte dans le cadre de l'actuel art. 9 de la Proposition

figure en en effet celui que le système porte atteinte aux droits fondamentaux des individus, une atteinte qui peut évidemment avoir lieu à bien des niveaux, la protection des données étant l'un d'entre eux. On peut toutefois se demander si, en lieu et place d'intégrer l'analyse algorithmique prévue à l'art. 9 de la Proposition au sein de l'analyse d'impact prévue par l'art. 35 RGPD (et désormais de l'art. 22 LPD), ce n'est pas cette dernière qui devrait être intégrée à l'analyse algorithmique prévue par l'art. 9 de la Proposition, puisque cette analyse couvre à mon sens un éventail de risques beaucoup plus large.

S'ajoute à la mise sur pied de cette analyse et de système de gestion des risques en découlant, un certain nombre de dispositions aux art. 10 et suivants de la Proposition des obligations renforçant cette gouvernance en ce qui a trait à (1) la gouvernance autour des données utilisées, (2) l'établissement d'une documentation technique, (3) l'enregistrement dans un registre central, (4) le devoir d'information et de transparence dus aux utilisateurs, (5) le contrôle humain ainsi que des exigences autour de la (6) l'exactitude, la robustesse et la cybersécurité.

Formellement, force est dès lors d'admettre que la Proposition de Règlement couvre assez largement les préoccupations relatives à la protection des données en imposant un cadre rigoureux au déploiement à tout le moins de ces systèmes à « risques élevés ». On ne saurait cependant voir dans cette couverture théorique une garantie pratique du respect de ces exigences.

Sans entrer dans les détails, la Proposition, qui est encore en cours d'examen, n'est en effet pas sans susciter quelques interrogations quant à la gouvernance nécessaire pour assurer sa mise en œuvre et son respect. Compte tenu du fait que cet examen risque fort d'être un autocontrôle, est-il sérieusement imaginable à ce jour d'avoir une agence par pays et une agence européenne à même de s'assurer du respect de l'ensemble de ces exigences, ne serait-ce que compte tenu de la multitude d'industries au sein desquelles ces outils peuvent être déployés, aux spécificités si différentes les unes des autres ? On peut en douter.

B. L'approche américaine

À ce jour, l'État fédéral n'a pas encore jugé utile de légiférer en la matière, et même si le parti Démocrate appelle désormais de ces vœux une réglementation en la matière, la pratique américaine consiste plutôt à laisser le soin aux différentes agences d'aborder certaines problématiques spécifiques les concernant au travers de recommandations ou de directives, que l'on pense par exemple au *US Copyright Office* ou à la *FTC*. Le risque existe donc que les différents États passent eux-mêmes à l'offensive en adoptant leur propre législa-

tion, tels l'État de la Californie ou celui du Colorado sur certains points spécifiques, avec pour conséquence un éventuel *patchwork* auquel les entreprises seraient soumises au gré de leurs activités ce qui, est-il besoin de le dire, n'apparaît guère optimal ; voilà donc une affaire à suivre.

Toujours est-il que l'appréciation des risques liés à la mise en œuvre de systèmes d'intelligence artificielle n'est évidemment pas absente des débats aux États-Unis, bien au contraire, puisque le *National Institute of Standards and Technology* a adopté le 26 janvier 2023 un document intitulé *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Ce document fournit un cadre désormais considéré comme un standard aux États-Unis et dont la prise en compte dépasse largement les frontières. Il s'articule autour de quatre piliers – ici largement résumés – que sont :

- « *Governance* » : la gouvernance implique la mise en place tant des ressources nécessaires que des processus tenant compte à la fois de l'environnement réglementaire et des risques liés au développement ainsi qu'à la mise en œuvre de ces systèmes tout au long de leur cycle de vie. Cette gouvernance s'applique à l'ensemble du cadre envisagé par le NIST AI RMF 1.0, et fait donc partie intégrante des trois autres piliers.
- « *Mapping* » : le déploiement de ces systèmes au sein d'une entreprise nécessite de s'assurer qu'ils s'intègrent à la stratégie de l'entreprise, que leur finalité et leurs objectifs sont bien définis, leurs limites et risques bien compris et que tous ces points sont dûment documentés, le tout s'insérant dans une analyse guidée par une approche coûts/bénéfices.
- « *Measure* » : suite à cette catégorisation des risques, il convient de les mesurer au travers de méthodologies adéquates (question évidemment loin d'être anodine…). L'analyse doit porter sur les éléments permettant de s'assure que le système est « digne de confiance » (« *trustworthy AI* »), ce qui intègre les aspects sécuritaires et de protection de la vie privée autour d'un audit de l'algorithme. Un tel contrôle doit avoir lieu non seulement au moment du déploiement, mais également lors de la vie de chacun de ces systèmes. L'un des objectifs poursuivis par cette phase consiste à satisfaire aux exigences de transparence (« *transparency* ») et d'explicabilité (« *explainability* ») de l'algorithme, même si l'effet « *black box* » rend un tel objectif potentiellement difficile à atteindre, tout dépendant au final du niveau d'exigences que l'on souhaite atteindre au travers de ces concepts.
- « *Manage* » : l'aspect gestion exige la mise en place des ressources nécessaires pour permettre la bonne exécution des aspects susmentionnés, leur documentation et la gestion des incidents susceptibles de survenir.

Le cadre proposé par le NIST AI RMF constitue de prime abord un cadre utile permettant la mise en œuvre de nombreuses exigences posées aux articles 9 et suivants de la Proposition de Règlement susmentionnée.

Ce n'est toutefois, et de loin, pas le seul référentiel possible.

III. Standards et gouvernance : le cœur du problème ?

Les standards ayant trait aux systèmes d'intelligence artificielle se multiplient, au point qu'il est aisé pour le néophyte de s'y perdre.

Pour bon nombre d'entre eux, ces standards sont élaborés au sein du sous-comité 42 du « *ISO/IEC JTC 1* », soit le comité technique commun à l'*International Organisation for Standardization (ISO)* et l'*International Electrotechnical Commission (IEC)*. À ce jour, le SC 42 a publié 17 standards, et 30 sont en développement.

Parmi ces standards, on mentionnera en particulier ISO/IEC 23894 :2023 en matière de gestion des risques, ISO/IEC 38507 : 2022 en matière de gouvernance, ainsi que les normes ISO/IEC 5259–5 concernant la qualité des données, ISO 42005 concernant l'analyse d'impact en matière d'intelligence artificielle ou encore ISO/IEC 6254 concernant l'explicabilité des systèmes.

L'environnement est donc riche. Se pose cependant la question de savoir si ces standards reflètent un large consensus et, à ce titre, de savoir quel est le processus d'élaboration de ces standards ; si le processus législatif est connu, celui des standards l'est moins, et il semblerait que leur contenu dépende pour l'essentiel du nombre de personnes d'un pays donné qui se trouvent à la table, avec une prédominance de certains acteurs… de là à dire qu'au final les absents ont toujours tort, il n'y a qu'un pas.

Au final, on constate que le cadre réglementaire se met en place, soutenu par des standards qui devraient en principe largement tenir compte des exigences posées en matière de protection des données et fournir une aide utile aux entreprises pour les aider dans la gouvernance de tels systèmes.

À mon sens, deux risques majeurs existent néanmoins à ce stade. Tout d'abord, celui d'une approche éclatée et fragmentée de standards aux notions revêtant des acceptions différentes et ne reflétant qu'un consensus au fond discutable, encadrés par des réglementations fort diverses d'un pays ou d'une région du monde à l'autre. Ensuite, compte tenu de

l'asymétrie d'informations qui existe aujourd'hui en matière d'intelligence artificielle entre les véritables experts et… le reste du monde, il existe un risque évident de ne voir aucune agence gouvernementale ou non gouvernementale à même d'assurer le contrôle du respect de ces exigences, exigences de surcroît applicables à des industries les plus diverses nécessitant des compétences impossibles à réunir au sein d'une seule et même agence. Il me semble donc que l'éducation et l'acquisition des connaissances sont des composantes essentielles pour permettre une bonne gouvernance de ces systèmes.

Poser un cadre est utile ; en permettre le contrôle et le respect est une nécessité qui, à ce jour, me semble loin d'être assurée. Affaire à suivre.

Proposition de citation : Philippe Gilliéron, Protection des données et intelligence artificielle : une gouvernance indispensable, 23 octobre 2023 *in* www.swissprivacy.law/259