

# Privacy-Enhancing Technologies et machine learning

## Une analyse par le prisme du droit de la protection des données

IAGO BAUMANN\*

MOTS CLEFS	Artificial Intelligence – Machine Learning – Data protection – Privacy-by-Design – Privacy-enhancing technologies (PETs) – Homomorphic Encryption – Federated Learning – Differential Privacy
ZUSAMMENFASSUNG	Der Beitrag untersucht die wichtigsten Datenschutztechnologien (Privacy-Enhancing Technologies, PETs) im Bereich des maschinellen Lernens – Anonymisierung, Pseudonymisierung, differenzielle Privatsphäre, föderiertes Lernen und homomorphe Verschlüsselung – und bewertet deren Einfluss auf den Datenschutz. Durch die Verbindung technischer und juristischer Perspektiven wird gezeigt, wie diese Werkzeuge zur Einhaltung des DSG und der DSGVO beitragen können, insbesondere im Sinne des Prinzips des Datenschutzes durch Technikgestaltung. Auch wenn klare <i>de lege lata</i> -Antworten schwer zu fassen sind, hebt die Studie konkrete Lösungen für Verantwortliche hervor und wirft zugleich grundlegende Fragen zur rechtlichen Regulierung der KI-Entwicklung auf.
RÉSUMÉ	Cet article examine les principales technologies de préservation de la vie privée (PETs) appliquées au machine learning – anonymisation, pseudonymisation, confidentialité différentielle, apprentissage fédéré, chiffrement homomorphe – et évalue leur impact sur la protection des données. En croisant approche technique et analyse juridique, il montre comment ces outils peuvent soutenir la conformité à la LPD et au RGPD, notamment via le principe de protection des données dès la conception. S'il reste difficile de donner des réponses tranchées <i>de lege lata</i> , l'étude met en lumière des solutions concrètes à la disposition des responsables de traitement, tout en soulevant les débats à mener pour encadrer juridiquement le développement de l'IA.
ABSTRACT	This article examines the main privacy-enhancing technologies (PETs) applied to machine learning – anonymisation, pseudonymisation, differential privacy, federated learning, and homomorphic encryption – and assesses their impact on data protection. By combining a technical approach with legal analysis, it shows how these tools can support compliance with the Swiss DPA and the GDPR, particularly through the principle of data protection by design. While it remains difficult to provide definitive answers <i>de lege lata</i> , the study highlights concrete solutions available to data controllers, while also raising key debates on the legal framework needed to govern the development of AI.

## I. Introduction

L'intelligence artificielle (ci-après : IA) est devenue incontournable depuis quelques années, perçue comme une révolution technologique capable d'augmenter l'efficacité et de réduire les coûts dans tous les domaines. Il s'agit généralement de systèmes basés sur du *machine learning* (ou apprentissage automatique), une technique visant à trouver des tendances au sein de grands volumes de données, dont souvent des données personnelles au sens de la loi sur la protection des données<sup>1</sup> (ci-après : LPD) suisse

ou du Règlement général sur la protection des données européen<sup>2</sup> (ci-après : RGPD). Cette technologie peut sembler, de par son fonctionnement et son appétit en données notamment personnelles<sup>3</sup>, difficilement compatible avec

\* IAGO BAUMANN, Doctorant en droit à l'Université de Neuchâtel, membre du LexTech Institute, MLaw.

Cette contribution est publiée sous une licence Creative Commons. DOI de cet article: 10.3256/978-3-03929-084-0\_02.

<sup>1</sup> Loi fédérale sur la protection des données du 25 septembre 2020 (LPD ; RS 235.1), art. 5 let. a.

<sup>2</sup> Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données, RGPD), art. 4(1).

<sup>3</sup> Notons qu'un modèle ne traitant que de données non personnelles ne soulève évidemment pas ce type de questions. On peut penser, à titre d'exemple, au logiciel NVIDIA DLSS, qui est une technologie d'*upscaling* basée sur l'intelligence artificielle, utilisée dans les jeux vidéo pour améliorer les performances graphiques et la fluidité en générant des images de haute qualité à partir de résolutions inférieures. Il fonctionne grâce à un modèle entraîné sur des images de jeux vidéo et ne

le droit de la protection des données<sup>4</sup>, ce qui pourrait encourager à chercher des solutions pour atténuer les conséquences<sup>5</sup>.

L'une des solutions est de recourir à des technologies qui permettent de forcer la conformité. Une sorte de contrôle de la technique par la technique. Ceci s'inscrit dans l'idée de la protection des données dès la conception. En effet, cette notion, présente tant dans la LPD que dans le RGPD, impose au responsable du traitement de données de tenir compte des principes de la protection des données personnelles dès la conception du traitement, notamment en mettant en place des mesures techniques et organisationnelles pour en garantir le respect<sup>6</sup>. Les *Privacy-Enhancing Technologies* (ci-après : PETs) sont précisément des mesures techniques visant à préserver la protection des données dans le contexte qui nous intéresse<sup>7</sup>. Elles jouent également un rôle déterminant en matière de sécurité des données<sup>8</sup>, aidant à préserver la confidentialité, l'intégrité, la disponibilité et la résilience des systèmes et des services traitant des données personnelles<sup>9</sup>. Nous verrons dans quelle mesure et par quels biais<sup>10</sup>.

---

traite aucune donnée personnelle des utilisateurs. Le droit de la protection des données ne s'y applique donc pas.

<sup>4</sup> BORIS P. PAAL, Spannungsverhältnis von KI und Datenschutzrecht, in : Markus Kaulartz/Tom Braegelmann (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 427 ss, 427 s., N 1 ss.

<sup>5</sup> Nous traiterons de cette potentielle incompatibilité entre l'IA et le droit de la protection des données (II C).

<sup>6</sup> Art. 7 al. 1 et 2 LPD ; art. 25(1) RGPD.

<sup>7</sup> LEE A. BYGRAVE, *The EU General Data Protection Regulation (GDPR)*, A Commentary, Oxford 2020, art. 25 573.

<sup>8</sup> Bien que la sécurité des données, couverte respectivement aux art. 8 LPD et 32 RGPD, fasse partie des principes de la LPD et du RGPD une distinction est souvent faite entre ce principe et la protection des données. En ce sens, voir le message du Conseil fédéral, Message du 15 septembre 2017 concernant la loi fédérale sur la révision totale de la loi fédérale sur la protection des données et sur la modification d'autres lois fédérales, FF 2017 6565 (ci-après : Message LPD), 6650 : « Il existe une interaction entre la protection des données et leur sécurité, mais ces deux aspects doivent être traités séparément. La protection des données relève de la protection de la personnalité de l'individu. Quant à la sécurité des données, elle vise généralement les données présentes chez un responsable du traitement ou chez un sous-traitant et englobe le cadre organisationnel et technique général du traitement des données. Par conséquent, la protection de l'individu n'est possible que si des mesures techniques générales ont été prises pour la sécurité des données le concernant. » Ce débat reste cependant très théorique, et ne change en rien le fait qu'une violation de la sécurité des données puisse être sanctionnée par le biais d'une loi de protection des données.

<sup>9</sup> Art. 8 LPD et 32(1)(b) RGPD.

<sup>10</sup> Section IV, en particulier A-C.

L'objectif de cet article est de s'intéresser à ces diverses mesures techniques, en observant les plus prometteuses dans le monde de la recherche en informatique, tout en s'interrogeant sur les différences que peuvent impliquer leur implémentation du point de vue du droit de la protection des données.

## II. Les frictions entre le droit de la protection des données et l'intelligence artificielle

### A. Le cycle de vie d'un système de machine learning

La conception et l'utilisation d'un modèle de *machine learning* suit généralement certaines étapes typiques. Le développeur commence par conceptualiser le problème qu'il tente de résoudre afin de pouvoir planifier les étapes suivantes du développement. Il sélectionne notamment le modèle semblant le plus adapté à la tâche, voire en développe un de sa propre facture. S'ensuit la phase de préparation des données qui comprend d'abord la collecte, puis la préparation des données (on parle de *preprocessing*)<sup>11</sup>. On essaie notamment à ce stade de limiter le « bruit »<sup>12</sup> dans les données. Une fois les données prétraitées, le développeur procède à l'entraînement du modèle d'IA en l'exposant à ces données. Il teste ensuite la qualité des sorties du modèle par rapport à la problématique à résoudre définie auparavant. Cette phase est généralement

---

<sup>11</sup> Le *preprocessing*, ou prétraitement des données, en *machine learning* comprend plusieurs étapes : le nettoyage (suppression des valeurs manquantes, anomalies et doublons), l'agrégation (regroupement de données issues de sources multiples), l'encodage des variables catégoriques, la mise à l'échelle (normalisation ou standardisation) et la réduction de dimensionnalité pour éliminer les redondances et éviter que certains types de données soient sur- ou sous-représentés. Enfin, les données sont divisées en ensembles d'entraînement, de validation et de test afin d'assurer une évaluation rigoureuse du modèle.

<sup>12</sup> En *machine learning*, le bruit désigne l'ensemble des variations aléatoires ou des erreurs présentes dans les données, susceptibles de perturber la modélisation et d'affecter la performance des algorithmes. Ce bruit peut être d'origine diverse : erreurs de mesure, imprécisions dans l'annotation des données, fluctuations naturelles des phénomènes observés ou encore artefacts introduits par le processus de collecte. S'il est trop important, il peut entraîner un surajustement (*overfitting*), où le modèle apprend des détails non généralisables au détriment des tendances sous-jacentes. Toutefois, dans certains contextes, l'ajout contrôlé de bruit est utilisé pour améliorer la robustesse des modèles ou préserver la confidentialité des données en limitant leur réidentification, comme nous allons le voir dans cette contribution.

itérative : le modèle est entraîné, puis validé, puis testé, puis réentraîné, puis validé, puis retesté et ainsi de suite, jusqu'à produire un résultat parvenant à satisfaire les attentes du développeur. Lorsque le modèle est considéré comme prêt, vient la phase de déploiement, afin de rendre le modèle utilisable par des tiers ou par la personne même qui l'a créé. Une fois déployé, le développeur continue de surveiller le modèle et peut parfois devoir procéder à des opérations de maintenance du système. C'est lors de cette phase que les utilisateurs interagissent avec le modèle, celui-ci devenant un outil dans lequel ils entrent des données et dont ils espèrent obtenir une sortie leur étant utile, relativement au cas d'usage<sup>13</sup>.

## B. Les types de *machine learning* et les cas d'usage

Il existe diverses techniques de *machine learning*, généralement choisies en fonction de la finalité du modèle. On peut notamment citer le *supervised learning* (apprentissage supervisé), qui repose sur l'entraînement à partir de données annotées (préparées) ; l'*unsupervised learning* (apprentissage non supervisé), qui cherche des *patterns* dans des données non annotées ; le *reinforcement learning* (apprentissage renforcé), qui optimise les décisions par un processus d'essais-erreurs guidé par des récompenses ; ou le *deep learning* (apprentissage profond), qui utilise des réseaux de neurones profonds pour modéliser des relations complexes<sup>14</sup>. Ces modèles peuvent être appliqués à un large éventail de cas, l'IA n'étant pas dépendante d'un champ en particulier, mais éminemment transversale. On trouve des systèmes de *machine learning* dans la santé, la finance, les transports, l'éducation, l'industrie, l'énergie ou l'agriculture. Les *Large-Language Models* (LLM) comme ChatGPT en sont encore un exemple.

## C. La collision entre le *machine learning* et la protection des données

La technologie de *machine learning*, lorsqu'elle est appliquée à des cas comprenant des données personnelles, tend naturellement à des frictions, voire des violations de dispositions du droit de la protection des données.

Les grands principes de la protection des données sont notamment la conformité à la finalité, qui comprend la

minimisation des données traitées et la limitation de la durée du traitement<sup>15</sup>, l'exactitude des données<sup>16</sup>, ainsi que le consentement éclairé de la personne concernée par le traitement<sup>17</sup>. Viennent s'ajouter à ces principes les droits des personnes concernées à être informées du traitement<sup>18</sup>, le droit d'accéder aux données traitées les concernant<sup>19</sup>, ainsi que le droit de les rectifier ou de les supprimer<sup>20</sup>. Notons aussi que le droit de la protection des données distingue divers types de données, dont les données sensibles telles que, par exemple, les données sur la santé, la confession, l'orientation sexuelle ou les opinions politiques, qui doivent être protégées le plus fermement. Rappelons enfin la notion de profilage adressée par la législation, qui impose des conditions<sup>21</sup> très marquées dans les cas où un traitement de données dresse un profil de personnalité, permettant par exemple de prédire des habitudes de consommation ou des préférences.

Plaçons ces principes et ces droits en perspective avec les aspects caractéristiques du *machine learning*, afin de mettre en lumière une apparente incompatibilité. Le *machine learning* nécessite de très larges volumes de données pour pouvoir être mis en place<sup>22</sup>. Une fois le modèle entraîné, les données peuvent rester comprises dans celui-ci et donc continuer d'être traitées tout au long de son cycle de vie. Certains modèles (notamment ceux à base de *deep learning*) ont pour but de trouver des schémas et des corrélations à partir de données sans que les développeurs sachent ce qui doit être trouvé, rendant impossible de définir une finalité en amont, ou alors celle-ci se doit d'être très vague<sup>23</sup>. Certains modèles infèrent des données à partir des données d'entraînement, or lorsque ces données d'entraînement sont relatives à des personnes, le résultat de l'inférence constitue également une donnée person-

<sup>15</sup> Art. 6 al. 3 s. LPD et 5(1)(b,c,e) RGPD.

<sup>16</sup> Art. 6 al. 5 LPD et 5(1)(d)RGPD.

<sup>17</sup> Art. 6 al. 6 s. LPD et 6(1)(a)RGPD.

<sup>18</sup> Art. 19 LPD et 13 s. RGPD.

<sup>19</sup> Art. 25 LPD et 15 RGPD. FÉLISE ROUILLE/ASTRID EPINEY, Le droit d'accès à ses données personnelles, in : Sylvain Métille (éd.), *Le droit d'accès*, Berne 2021, 2 s.

<sup>20</sup> Art. 32 LPD et 16 s. RGPD.

<sup>21</sup> Notamment art. 5 let. f et g, 6 al. 7 LPD ; notamment art. 4 et 22 RGPD.

<sup>22</sup> SOPHIE EVERARTS DE VELP, Big Data dans l'IA et principe de minimisation : Défis et risques, in : Hervé Jacquemin (éd.), *Time to reshape the digital society*, Bruxelles 2021, 293 ; FRANÇOIS CHARLET, Protection des données en entreprise, Bâle 2023, 323 s.

<sup>23</sup> TINA GAUSLING, Datenschutzrechtliche Informationspflichten, in : Markus Kaulartz/Tom Braegelmann (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 379 ss, 382, N 13 ss ; CHARLET (n. 22), 325.

<sup>13</sup> Pour une décomposition des étapes plus détaillée, voir BRENDAN QUINN, *Data Protection Implementation Guide, A Legal, Risk and Technology Framework for the GDPR*, Croydon 2021, 199 s.

<sup>14</sup> QUINN (n. 13), 196 ss.

nelle et il n'est pas évident d'admettre leur exactitude<sup>24</sup>. Lorsque la finalité et la durée du traitement sont difficiles à définir, peut-on réellement considérer que la personne concernée par des données comprises dans le *dataset* d'entraînement soit en mesure de donner un consentement éclairé<sup>25</sup> ?

À l'ère du *Big Data* et de l'extraction de données (*data scraping* et *data mining*) à outrance, comment informer toutes les personnes dont les données ont été collectées pour entraîner un modèle<sup>26</sup> et dans quelle mesure les informer des détails du traitement<sup>27</sup> ? Lorsque la phase d'entraînement est terminée, compte tenu du caractère de quasi-boîte noire<sup>28</sup> du modèle entraîné, il est parfois très difficile de déterminer si une donnée en particulier est comprise dans les poids de celui-ci et encore plus difficile voire impossible de modifier ou de supprimer une telle donnée<sup>29</sup>. Bon nombre de systèmes de *machine learning* traitent des données sensibles ou ont recours à du profilage<sup>30</sup>. On peut, par exemple, citer les modèles de reconnaissance faciale (données biométriques), les mo-

dèles radiologiques (données de santé), les algorithmes de suggestions commerciales (profilage des habitudes de consommation) ou de contenu (profilage des goûts artistiques ou des opinions politiques)<sup>31</sup>.

Toutes ces questions, dont les traits ont volontairement été grossis et présentés brièvement, mettent en évidence le conflit profond entre la protection des données et l'IA<sup>32</sup>, sans pour autant diaboliser cette dernière, bien au contraire. L'objectif des PETs est précisément de parvenir à des approches de *machine learning* plus respectueuses des données personnelles, tout en favorisant une poursuite des développements techniques afin de profiter des nombreuses opportunités offertes par cette révolution technologique.

### III. Les aspects techniques des PETs en matière de *machine learning*

Cette section prolonge la perspective technique déjà amorcée, notamment par la présentation des bases du *machine learning*. Il s'agit ici d'adopter une approche similaire pour les PETs, avant d'en examiner les implications juridiques.

Nous faisons donc le choix assumé d'une incursion technique, parfois éloignée du raisonnement juridique classique. Mais dans un domaine aussi profondément interdisciplinaire, il serait illusoire de penser le droit sans comprendre, même dans les grandes lignes, les mécanismes techniques à l'œuvre. Se limiter à une analyse strictement juridique risquerait de produire un raisonnement abstrait, déconnecté des réalités concrètes, et donc d'une portée limitée.

Le droit de la protection des données repose en effet de plus en plus sur des notions dont la portée dépend directement des solutions technologiques mobilisées. Comprendre ces dernières, même de manière simplifiée, permet d'évaluer plus justement les marges de manœuvre normatives, les enjeux de qualification juridique, ou encore les effets concrets des obligations pesant sur les res-

<sup>24</sup> CHARLET (n. 22), 324 ; AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS (AEPD), Artificial Intelligence: accuracy principle in the processing activity, 31 mai 2023 ; EUROPEAN DATA PROTECTION SUPERVISOR (EDPS), Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI systems, 15.

<sup>25</sup> Voir notamment FLORIAN HÉMONT/MARINE GOUT, Consentement résigné : en finir avec le *Privacy Paradox*, in : Alexandra Bensamoun/Maryline Boizard/Sandrine Turgis (éd.), Le profilage en ligne : entre libéralisme et régulation, 21 ss, 33 ss.

<sup>26</sup> Sur la question du *scraping* sur les réseaux sociaux, qui est l'une des manières courantes de collecter des données à des fins d'entraînement de modèles, notamment de LLM, voir CATHERINE ALTOBELLI et al., To Scrape or Not to Scrape? The Lawfulness of Social Media Crawling under the GDPR, in : Jean Herveg (éd.), Deep Diving into Data Protection, Bruxelles 2021, 151 ss, 160 ss.

<sup>27</sup> GAUSLING (n. 23), 379 ss.

<sup>28</sup> En *machine learning*, un modèle est qualifié de « boîte noire » lorsque son fonctionnement interne est difficile à comprendre. Bien que l'on puisse observer ses entrées et sorties, les calculs et les transformations effectués pour arriver au résultat restent opaques, rendant l'interprétation et l'explication des décisions complexes.

<sup>29</sup> Problématique connue sous le nom du *Unlearning Problem*, ou comment faire désapprendre une information à un algorithme d'IA. Voir notamment CHARLET (n. 22), 326 ; KAIRAN ZHAO et al., What makes unlearning hard and what to do about it, NeurIPS 2024 ; JAYDEEP BORKAR, What can we learn from Data Leakage and Unlearning for Law?, the first GenLaw workshop at ICML'23, Hawai'i 2023 ; BJØRN ASLAK JULIUSSEN/JON PETER RUI/DAG JOHANSEN, Algorithms that forget: Machine unlearning and the right to erasure, Computer Law & Security Review, vol. 51, Novembre 2023.

<sup>30</sup> Pour de larges approfondissements concernant cette problématique, voir OLIVIER HEUBERGER, Profiling im Persönlich-

keits- und Datenschutzrecht der Schweiz, thèse Lucerne 2020, en particulier 125 ss.

<sup>31</sup> Voir notamment TRISTAN ALLARD et al., Ouvrir la boîte noire des algorithmes de personnalisation, in : Alexandra Bensamoun/Maryline Boizard/Sandrine Turgis (éd.), Le profilage en ligne : entre libéralisme et régulation, 211 ss, 214 ss ; QUINN (n. 13), 201 s.

<sup>32</sup> PAAL (n. 4), 427 s., N 1 ss. Dans le même sens, voir notamment SÉBASTIEN FANTI, Protection des données informatiques, in : Jean-Philippe Dunand/Pascal Mahon (éd.), La protection des données dans les relations de travail, Genève/Zurich/Bâle 2017, 229 ss, 261 s.

posables de traitement. Ce détour par la technique n'est donc pas accessoire : il constitue une condition préalable à une analyse juridique solide et pertinente.

### A. La notion de PETs

Les PETs sont des techniques conçues pour protéger les données personnelles et/ou sensibles tout en permettant leur utilisation. Elles minimisent ou éliminent les risques pour la vie privée en limitant l'accès, en anonymisant, en chiffrant ou en réduisant l'accès d'un responsable de traitement sur les données, tout en essayant de ne pas compromettre les objectifs des analyses ou des services qui les exploitent. Les VPN (réseaux privés virtuels), le chiffrement de bout en bout<sup>33</sup> (que l'on peut retrouver dans des messageries comme Whatsapp ou Signal) ou le contrôle des permissions des applications (proposé notamment sur Android et iOS et permettant à l'utilisateur d'interdire l'accès par une application particulière à certaines données comme les photos) sont tant d'exemples de ces technologies. Les PETs sont évidemment très dépendantes du service en question et doivent par conséquent être adaptées aux divers cas d'usage.

Les PETs les plus faciles à comprendre dans leur conception et, partant, venant en premier à l'esprit sont les notions de pseudonymisation et d'anonymisation, dont nous présenterons les avantages et les limites (B). Les technologies, plus révolutionnaires et plus prometteuses, que nous traiterons ensuite (C à E) sont la confidentialité différentielle (*differential privacy*, une version évoluée de l'idée d'anonymisation), l'apprentissage fédéré (qui décentralise les différentes phases du cycle de vie d'un système de *machine learning*, interdisant le monopole des données à un seul acteur), le chiffrement homomorphe (*homomorphic encryption*, qui se base sur des concepts de cryptographie avancée). Notons également l'existence d'autres PETs comme le *zero-knowledge machine learning* (zkML)<sup>34</sup>, l'entraînement avec bruit (*Noise Injection in Training*, NIT, différent de la confidentialité différentielle car ce premier s'applique sur les poids et gradients du modèle, tandis que la seconde place du bruit sur les données d'entraînement ou potentiellement sur l'inférence obtenue en bout de chaîne), les environnements d'exécution sécurisés

(*Trusted Execution Environments*, TEE, par exemple faire les calculs sur une partie du processeur, coupée/protégée du reste et du monde extérieur)<sup>35</sup>, ou les données synthétiques (*synthetic data*, données artificielles, créées par des modèles génératifs et ressemblant aux données réelles, mais sans contenir d'informations sensibles). Nous nous limitons à détailler les premières solutions exposées, car, si l'on en croit le volume de recherche respectivement effectuée sur chacune de ces PETs en matière d'IA, elles semblent aujourd'hui les plus prometteuses, sans pour autant écarter la potentielle portée future de ces dernières.

Ces différentes PETs sont généralement combinées dans des flux de travail (*workflows*) et sont donc compatibles et cumulables entre elles, rendant le modèle résultant d'autant plus respectueux de la protection des données. Notons toutefois que, malgré les aspects très positifs de l'implémentation de PETs du point de vue de la protection et de la sécurité des données, l'efficacité du modèle peut s'en retrouver affectée<sup>36</sup>.

### B. La pseudonymisation, l'anonymisation et le risque de ré-identification

Il serait intéressant de pouvoir dépersonnaliser les données personnelles, d'une part, pour éviter la réalisation des risques pour la vie privée des personnes concernées que cherche à protéger le droit de la protection des données<sup>37</sup> et, d'autre part, pour pouvoir utiliser les données avec plus de liberté. Tant la pseudonymisation que l'anonymisation poursuivent cet objectif. Les questions entourant ces deux techniques, ainsi que leurs limites ont été et continuent d'être extensivement discutées en doctrine<sup>38</sup>.

La pseudonymisation est l'opération consistant à altérer les données afin qu'elles « ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel

<sup>33</sup> DURMUS ERDEM, Privacy by Design und Privacy by Default: Alles nur eine Frage der Einstellung?, DSB 2024, 236 ss, 238.

<sup>34</sup> Qui a le potentiel de devenir révolutionnaire en la matière, mais dont la recherche se trouve encore à un stade embryonnaire. Pour une analyse du potentiel du zkML d'un point de vue juridique, en particulier en protection des données, voir IAGO BAUMANN, zkML and Data Protection, LexTech Institute, blog mai 2025.

<sup>35</sup> MANON KNOCKAERT et al., Privacy-by-Design in Intelligent Infrastructures, in : Jean Herveg (éd.), Deep Diving into Data Protection, Bruxelles 2021, 309 ss, 332 ; LIV D'ALBERTI/EVAN GRONBERG/JOSEPH KOVBA, Privacy-Enhancing Technologies for Artificial Intelligence-Enabled Systems, IWSPA '24, Porto 2024, 3.1, 4.

<sup>36</sup> Nous développons brièvement cette question ci-après, III F.

<sup>37</sup> Voir notamment art. 1 LPD ; CR LPD-COTTIER, art. 1 N 15 ss ; art. 1 RGPD ; HIELKE HIJMANS, The EU General Data Protection Regulation (GDPR), A Commentary, art. 1 50 s.

<sup>38</sup> Voir notamment ALEXANDRE JOTTERAND, Personal Data or Anonymous Data: where to draw the lines (and why)?, Jusletter 15 août 2022, et références.

ne sont pas attribuées à une personne physique identifiée ou identifiable »<sup>39</sup>. La personne concernée reste ainsi identifiable, du moins pour les personnes disposant des bonnes informations<sup>40</sup>. Le pseudonyme peut être obtenu au simple hasard, en recourant à de la cryptographie<sup>41</sup>, ou en décidant d'une logique.

L'anonymisation, quant à elle, est une opération rendant la personne concernée non identifiable de manière irréversible<sup>42</sup>, et ce même pour le responsable du traitement<sup>43</sup>. On procède en supprimant les identifiants personnels, comme le nom, le numéro de téléphone, l'adresse, le numéro AVS, etc. L'objectif est de parvenir au stade où les données restantes dans la base de données ne soient plus suffisantes pour distinguer la personne se cachant derrière les données<sup>44</sup>.

On s'aperçoit que la notion d'identifiabilité est centrale<sup>45</sup>. Elle est très largement débattue en doctrine, à laquelle nous renvoyons pour des détails<sup>46</sup>.

Les limites à ces méthodes sont importantes. Outre le fait que les données perdent en détails, puisqu'elles sont amputées de certains de leurs éléments, ce qui est déjà une limite en soi, l'avènement du *Big Data* a mis en évidence la grande difficulté à effacer le lien entre une donnée et une personne<sup>47</sup>. On parle généralement du problème de ré-identification. En effet, par le biais de techniques avancées d'analyse de données, croisant des données anonymisées avec d'autres bases de données titanesques, il est possible de retrouver les personnes concernées, et ce de plus en plus facilement<sup>48</sup>. Un fameux exemple, vieux de plus de quinze ans, illustre cette facilité : dans les années 2000,

quatre-vingt-sept pour cent de tous les américains pouvaient être identifiés sur la seule base de leur code postal, leur genre et leur date de naissance<sup>49</sup>. Trois données pouvant sembler plutôt anonymes, ou du moins difficiles à lier directement à un individu, permettent en réalité de le faire pour une très large majorité de la population. Il semble évident qu'avec l'augmentation exponentielle des données produites, collectées et analysées depuis lors, la situation n'a fait qu'empirer. Certaines stratégies permettent de mitiger ce risque de ré-identification, mais jamais totalement<sup>50</sup>.

### C. La confidentialité différentielle

La confidentialité différentielle (*differential privacy*) vise à garantir qu'une information anonymisée ne puisse être attribuée à un individu précis qu'avec une probabilité maximale prédéfinie, même si un attaquant dispose des données originales des autres individus<sup>51</sup>. Elle repose sur le principe selon lequel l'ajout ou la suppression d'un enregistrement individuel doit entraîner une variation minimale dans la distribution des résultats produits par la fonction d'anonymisation, dépassant ainsi les approches classiques basées sur des blocs identiques de quasi-identifiants<sup>52</sup>.

Cette méthode consiste généralement à ajouter un bruit aléatoire aux données pour limiter les inférences précises sur leurs propriétés<sup>53</sup>. Elle est le plus souvent appliquée lors du prétraitement, après la collecte des données, afin de protéger les informations utilisées à l'entraînement, mais peut aussi intervenir à la phase d'inférence, en bruitant les réponses aux requêtes pour éviter les fuites de données sensibles<sup>54</sup>. Par exemple, des diagnostics comme « grippe » et « rhume » peuvent être regroupés sous une catégorie générique comme « infection », réduisant ainsi

<sup>39</sup> Art. 4(5) RGPD; CR LPD-MEIER/TSCHUMY, art. 5 N 29.

<sup>40</sup> QUINN (n. 13), 160.

<sup>41</sup> Les informations en clair peuvent être hashées, afin d'obtenir un identifiant illisible pour quiconque ne connaît pas la table de hachage, voir QUINN (n. 13), 30.

<sup>42</sup> Consid. 26 RGPD; CR LPD-MEIER/TSCHUMY, art. 5 N 27; WILLIAM PATRICK HISLAIRE, L'autodétermination informationnelle, thèse Fribourg 2023, 172, N 229 s.

<sup>43</sup> QUINN (n. 13), 30 s.

<sup>44</sup> HISLAIRE (n. 43), 172, N 230.

<sup>45</sup> RUNSHAN HU et al., Bridging Policy, Regulation and Practice?, in : Ronald Leenes et al. (éd.), *Data Protection and Privacy, The Age of Intelligent Machines*, Oxford 2017, 115 ss, 120.

<sup>46</sup> JOTTERAND (n. 38), N 11 ss; CR LPD-MEIER/TSCHUMY, art. 5 N 21 ss et références; DAVID ROSENTHAL, La nouvelle loi sur la protection des données, Jusletter 16 novembre 2020, N 19 ss; CHARLET (n. 22), 22 ss. En droit européen, voir notamment LEE A. BYGRAVE/LUCA TOSONI, *The EU General Data Protection Regulation (GDPR)*, A Commentary, art. 4(1) 110 s. et références.

<sup>47</sup> SUZANNE VERGNOLLE, L'effectivité de la protection des personnes par le droit des données à caractère personnel, thèse Panthéon-Assas 2020, Bruxelles 2022, 154, N 185.

<sup>48</sup> CR LPD-MEIER/TSCHUMY, art. 5 N 28.

<sup>49</sup> NATE ANDERSON, "Anonymized" data really isn't—and here's why not, blog 8 septembre 2009.

<sup>50</sup> EVERARTS DE VELD (n. 22), 295 ss; YVES DONZALLAZ, *Traité de droit médical*, vol. 2, Berne 2021, 2940, N 6182; NATHANAËL PASCAL, L'anonymisation, une fausse bonne idée?, 13 novembre 2024, in : [www.swissprivacy.law/323](http://www.swissprivacy.law/323).

<sup>51</sup> ERIK BUCHMANN, Anonymitätssmasse für Personendaten, *digma* 2011, 166 ss, 169.

<sup>52</sup> JOSEPH NEAR/DAVID DARAI/KAITLIN BOECKL, *Differential Privacy for Privacy-Preserving Data Analysis*, blog, 27 juillet 2020.

<sup>53</sup> LISA KÄDE/STEPHANIE VON MALTZAN, Algorithmen, die nicht vergessen – Sicherheitslücken in Machine-Learning-Modellen und deren Bedeutung für den Schutz der Daten und der Urheberrechte, *InTeR* 2020, 201 ss, 208.

<sup>54</sup> DAYONG YE et al., One Parameter Defense – Defending against Data Inference Attacks via Differential Privacy, 13 mars 2022.

les risques d'identification. Cette technique tient également compte des mises à jour des données et du degré de connaissance possible d'un attaquant. Bien qu'efficace pour garantir la confidentialité, elle peut entraîner une perte d'information significative selon la distribution des données. Des recherches sont en cours<sup>55</sup> pour concevoir des variantes moins contraignantes maintenant les bénéfices de confidentialité tout en réduisant l'impact sur la qualité des données, par exemple en limitant ce qu'un attaquant peut théoriquement apprendre.

L'application de la confidentialité différentielle suppose en général que le responsable du traitement détienne une base de données en clair (non chiffrée), sur laquelle il applique des transformations visant à masquer certaines informations<sup>56</sup>, dans la continuité des techniques d'anonymisation ou de pseudonymisation, mais de manière plus avancée.

#### D. L'apprentissage fédéré

L'apprentissage fédéré (ou *federated learning*) est une forme d'apprentissage collaboratif apparue au cours des dix dernières années et qui suscite un intérêt croissant. Contrairement au *machine learning* classique, où les données sont centralisées pour entraîner un modèle, cette approche inverse la logique : c'est le modèle qui est envoyé vers les données.

Concrètement, un modèle central est d'abord initialisé, fournissant une base commune à tous les participants. Ceux-ci sont ensuite sélectionnés selon des critères spécifiques (pertinence, diversité, etc.) et reçoivent le modèle avec un programme d'entraînement. Chacun l'entraîne localement sur ses propres données, sans jamais les transmettre. Seules les mises à jour du modèle sont ensuite renvoyées et agrégées par un serveur central pour enrichir le modèle global. Ce cycle est répété plusieurs fois afin d'optimiser le modèle final, tout en garantissant la confidentialité des données, qui restent localement stockées sur l'appareil du client, sans que le propriétaire du modèle global ne puisse voir les données en clair<sup>57</sup>.

L'un des grands avantages de cette approche est de permettre un apprentissage sur données brutes, sans recourir à des techniques comme le chiffrement ou l'ajout de bruit, souvent néfastes à la qualité –et donc à la performance – du modèle<sup>58</sup>.

#### E. Le chiffrement homomorphe

Le chiffrement homomorphe est une technique cryptographique qui permet de réaliser des calculs sur des données chiffrées, sans jamais les déchiffrer. Ainsi, un tiers (par exemple, l'hôte du modèle<sup>59</sup>) peut manipuler ou interroger ces données tout en garantissant que ni les données d'origine ni les résultats intermédiaires ne lui sont accessibles<sup>60</sup>.

Contrairement aux méthodes classiques où les données doivent être déchiffrées pour être traitées, ici les opéra-

---

des participants. De telles failles soulignent la nécessité d'intégrer des mécanismes complémentaires de protection, tels que la régularisation différentielle ou l'apprentissage fédéré avec confidentialité renforcée, afin de limiter les risques liés à la divulgation indirecte d'informations sensibles. Pour des détails concernant le fonctionnement technique général de l'apprentissage fédéré, voir notamment MICHAEL R. A. HUTH, *Federated Learning*, in : Markus Kaulartz/Tom Braegelmann (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 37 ss, 39 ss, N 13 ss ; MARCO SCHREYER et al., *A sum greater than its parts: collective artificial intelligence in auditing*, *Expert Focus* 4/24, 180 ss, 182 s. ; PETER KAIROUZ et al., *Advances and Open Problems in Federated Learning*, *Foundations and Trends in Machine Learning*, vol 4 Issue 1, 10 décembre 2019.

<sup>58</sup> MARKUS KAULARTZ, *Personenbezug von KI-Modellen*, in : Markus Kaulartz/Tom Braegelmann (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 462 ss, 470 s., N 26 ss.

<sup>59</sup> L'hôte d'un modèle désigne l'entité responsable de l'hébergement et de l'administration d'un modèle d'intelligence artificielle, incluant son architecture, ses paramètres et les ressources informatiques nécessaires à son exécution. Il peut s'agir d'un fournisseur de services cloud, d'une entreprise technologique ou d'une institution académique. Par exemple, OpenAI est l'hôte du modèle ChatGPT, qu'il déploie via une infrastructure dédiée permettant aux utilisateurs d'y accéder sans exposer directement les poids du modèle. En règle générale, l'hôte a accès aux données traitées par le modèle, car elles doivent être déchiffrées pour être exploitées. Toutefois, dans le cas du chiffrement homomorphe, les calculs sont effectués directement sur des données chiffrées, empêchant ainsi l'hôte d'accéder aux informations en clair, tant au niveau de la requête qu'au niveau du résultat de l'inférence.

<sup>60</sup> KNOCKAERT et al. (n. 35), 331 ; MICHAEL HEFERT/BENJAMIN LANGE/DOMINIK SPYCHALSKI, *Verschlüsselung in der Cloud*, *digma* 2019, 128 ss, 132 ; KONSTANTIN G. KOGOS et al., *Fully Homomorphic Encryption Schemes: the State of The Art*, *IEEE Xplore* 2017, 463 ss, 463 ss.

<sup>55</sup> Voir notamment Rachel Cummings et al., *Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment*, *Harvard Data Science Review*, 6(1), 16 janvier 2024.

<sup>56</sup> KNOCKAERT et al. (n. 35), 332.

<sup>57</sup> Toutefois, malgré les garanties de confidentialité offertes par cette approche, certaines vulnérabilités subsistent. Parmi elles, les attaques par inférence d'appartenance (*membership inference attacks*) constituent une menace notable. Ces attaques exploitent les mises à jour successives du modèle global afin d'inférer si des données spécifiques ont été utilisées lors de l'entraînement, ce qui peut compromettre la confidentialité

tions (comme des additions ou multiplications) sont effectuées directement sur les données chiffrées. Lors du déchiffrement final, le résultat est identique à celui qu'on aurait obtenu en effectuant les mêmes calculs sur les données en clair. Ce fonctionnement repose sur des propriétés mathématiques spécifiques liées à l'homomorphisme, qui permet la préservation de certaines opérations sous chiffrement<sup>61</sup>. Ceci assure donc que les données en question ne soient jamais déchiffrées au cours de l'opération et ne soient donc jamais visibles en clair par le détenteur du modèle réalisant l'opération<sup>62</sup>.

En pratique, bien que cette méthode puisse s'appliquer à la phase d'entraînement, elle reste très coûteuse en calcul, surtout sur de grands volumes de données<sup>63</sup>. En revanche, elle est particulièrement adaptée à la phase d'inférence, permettant aux utilisateurs d'interroger un modèle tout en protégeant la confidentialité de leurs données d'entrée comme de sortie.

## F. L'arbitrage entre performance et protection

Le principal frein à une adoption plus large et rapide des PETs dans le développement de systèmes de *machine learning* reste sans doute le coût en performances qu'elles impliquent<sup>64</sup>.

Les méthodes d'anonymisation, de pseudonymisation ou de confidentialité différentielle nécessitent un traitement complet de la base de données pour réduire les liens identifiants représentant une charge de travail importante. Plus la base est volumineuse et les opérations complexes, plus ce coût augmente.

L'apprentissage fédéré requiert des échanges fréquents entre les nœuds durant l'entraînement, mobilisant puissance de calcul, mémoire locale et large bande passante pour la transmission des paramètres. Il évite cependant la centralisation des données, elle-même coûteuse en ressources. Néanmoins, cette approche reste limitée par les contraintes techniques des machines décentralisées et la qualité de leur connexion<sup>65</sup>.

Le chiffrement homomorphe présente également des limites notables. Sa performance est freinée par des exigences computationnelles élevées, particulièrement pour les grands ensembles de données ou les calculs complexes, entraînant un surcoût significatif en traitement. De plus,

sa fonctionnalité reste limitée, incapable de prendre en charge tous les types de calculs. Enfin, la gestion des clés, essentielle pour sécuriser les données chiffrées, reste un défi en termes de protection et de contrôle des accès<sup>66</sup>.

En pratique, l'implémentation des PETs exige une expertise pointue<sup>67</sup>, aujourd'hui surtout présente dans le monde académique. Cela entraîne des coûts de recrutement non négligeables pour les entreprises souhaitant internaliser ces compétences, créant un obstacle supplémentaire<sup>68</sup>.

L'espoir réside dans les progrès de la recherche, qui pourraient rendre ces techniques plus efficaces et réduire le compromis entre performance et protection des données. Si les modèles futurs atteignent des performances similaires (voire supérieures) à coût égal ou inférieur, et que les PETs deviennent plus accessibles, renforcer la protection des données apparaîtra moins comme un sacrifice. Dans un contexte capitaliste et consumériste, la vie privée reste un argument secondaire pour la majorité des clients, souvent éclipsé par les fonctionnalités et performances de l'IA. Dès lors, les développeurs sont peu incités à concevoir des modèles véritablement respectueux des données.

## IV. Les conséquences juridiques en matière de protection des données en cas d'implémentation de PETs dans un système de *machine learning*

### A. L'inapplicabilité du droit de la protection des données

Le droit de la protection des données ne s'applique pas aux données anonymisées<sup>69</sup>. Si le lien d'identification entre la personne et la donnée a été irréversiblement supprimé, ou que les efforts devant être consentis pour le rétablir font qu'aucun intéressé ne s'y attachera, alors ni la LPD ni le RGPD ne s'appliquent<sup>70</sup>. La question de ce que sont des efforts dont le prix serait suffisamment dissuasif reste donc déterminante. Cette idée évolue au gré de l'état de la

<sup>61</sup> DONZALLAZ (n. 50), 2943, N 6189.

<sup>62</sup> D'ALIBERTI/GRONBERG/KOVBA (n. 35), 4 s., N 3.1.

<sup>63</sup> D'ALIBERTI/GRONBERG/KOVBA (n. 35), 5, N 3.4.

<sup>64</sup> ELIZABETH RENIERIS, Why PETs (privacy-enhancing technologies) may not always be our friends, blog, 29 avril 2021.

<sup>65</sup> Pour le paragraphe, KAIROUZ et al. (n. 57).

<sup>66</sup> Pour le paragraphe, WILFRED W. K. LIN, Challenges of Homomorphic encryption, avril 2023.

<sup>67</sup> INFORMATION COMMISSIONER'S OFFICE (ICO), Chapter 5: Privacy-enhancing technologies (PETs). Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance, septembre 2022, 5.

<sup>68</sup> THE ROYAL SOCIETY, From privacy to partnership – Policy report, 2023, 36 s.

<sup>69</sup> Art. 2 al. 1 et 5 let. a LPD ; art. 2(1) et 4(1), ainsi que consid. 26 RGPD.

<sup>70</sup> Message LPD (n. 8), 6639 s. ; consid. 26 RGPD. Voir aussi CR LPD-FANTI/STAEGGER, art. 8 N 104.

technique, mais aussi avec la valeur des données en question. Les données personnelles, notamment permettant le profilage d'individus semblent gagner toujours plus de valeur, ce qui pourrait rendre intéressant de consentir à des efforts très conséquents pour chercher à casser une anonymisation trop peu robuste. Comme nous l'avons vu, le risque de ré-identification est considérable<sup>71</sup> et on peut, partant, admettre qu'une large majorité des données qui pourraient être qualifiées d'anonymes ne le sont pas. Selon nous, elles doivent ainsi rester soumises au droit de la protection des données, et ce même en cas d'application des dernières techniques découvertes par la recherche, compte tenu de la difficulté d'éliminer absolument tout risque résiduel d'identification<sup>72</sup>. D'abord, aucune technologie présentée dans cette contribution n'est exempte de failles<sup>73</sup>. De plus, même si l'on pose l'hypothèse d'une PET parfaitement dé-identifiante ou anonymisante, et donc libératrice des obligations imposées par le droit de la protection des données, il faudrait encore s'interroger sur les risques résiduels. Les PETs peuvent avoir une tendance à donner une fausse impression de protection et de sécurité absolue<sup>74</sup>. Le responsable de traitement (qui n'en est plus un) serait potentiellement tenté de ne plus tenir compte d'aucun des préceptes portés par le droit de la protection des données, par exemple la minimisation ou l'exactitude des données, ce qui pourrait entraîner des résultats néfastes à une échelle plus large<sup>75</sup>, sans pour autant que les données d'une personne physique en particulier soient atteintes.

La question découlant de ceci est de déterminer le rôle réel du droit de la protection des données. Ce droit doit-il avoir pour seul but de protéger les données personnelles d'individus en particulier, ou devrait-il tenir compte de

considérations plus larges, à savoir que l'agglomération de données sur de grandes masses d'êtres humains, assistée d'outils permettant l'analyse de cet agglomérat, pourraient mener à des dérives pouvant influencer nos vies et nos démocraties, même sans que les données puissent être liées directement à quelqu'un ? Ces questions, hors du cadre de cet article, sont partiellement traitées dans des dispositions spécifiques du nouveau Règlement européen sur l'IA<sup>76</sup> en ce qui concerne les modèles d'IA à usage général comportant des risques systémiques pour la société.

Pour le cas des données pseudonymisées, étant identifiables aux yeux du responsable du traitement, elles constituent des données personnelles<sup>77</sup>. Le risque de ré-identification intervient également ici, même sans fuites des clés permettant normalement d'identifier la personne derrière le pseudonyme.<sup>78</sup> Notons que les données dont la tentative d'anonymisation n'a pas abouti à de réelles données anonymisées, deviennent en réalité une forme de données pseudonymisées.

Notons enfin qu'il n'est pas certain qu'une donnée présente dans la base d'entraînement puisse être retrouvée, en tant que telle, dans le modèle final. Cette question, qui touche à la persistance des données personnelles dans les paramètres appris, est fondamentale. Elle dépasse cependant le cadre du présent article et mériterait une analyse spécifique, tant ses implications pourraient remettre en question une partie des conclusions ici formulées. En effet, si l'on considérait qu'aucune donnée personnelle ne subsiste dans un modèle une fois l'entraînement terminé, la qualification juridique du traitement changerait sensiblement.

Cela étant dit, il apparaît clairement, à la lumière de nombreuses recherches empiriques<sup>79</sup>, que certaines données peuvent bel et bien être extraites ou inférées à partir de modèles entraînés – en particulier dans le cadre de techniques d'attaque ciblées. Dès lors, les risques d'identifiabilité ne peuvent être écartés. Nos réflexions conservent donc toute leur pertinence, en particulier dans la mesure où la notion d'identifiabilité demeure centrale et, comme nous l'avons vu (III B), sujette à interprétation et à débat.

<sup>71</sup> Section III B.

<sup>72</sup> Concernant le débat entre relativisme et absolutisme dans l'interprétation de l'identifiabilité d'une donnée, voir JOTTERAND (n. 38), en particulier N 11 ss.

<sup>73</sup> Nous avons largement discuté des limites de l'anonymisation et de la pseudonymisation. Pour la differential privacy, voir JOSEF DOMINGO-FERRER/DAVID SÁNCHEZ/ALBERTO BLANCO-JUSTICIA, *The Limits of Differential Privacy (and its Misuse in Data Release and Machine Learning)*, 4 novembre 2020 ; MONIKA VALKANOVA, *Trainieren von KI-Modellen*, in : Markus Kaulartz/Tom Braegelmann (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 336 ss, 351, N 52 s. Pour le federated learning, voir KAIROUZ et al. (n. 57).

<sup>74</sup> THE EUROPEAN COMMISSION DG JUSTICE, FREEDOM AND SECURITY (LONDON ECONOMICS), *Study on the economic benefits of privacy-enhancing technologies (PETs)*, Juillet 2010, 8.

<sup>75</sup> On peut penser à un pouvoir de manipulation accru sur les populations, grâce à l'accumulation par des acteurs (réseaux sociaux, État, publicitaires) de données comportementales sur un large nombre d'individus, par exemple.

<sup>76</sup> Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'intelligence artificielle, RIA).

<sup>77</sup> Voir notamment consid. 26 RGPD.

<sup>78</sup> LUCA TOSONI, *The EU General Data Protection Regulation (GDPR)*, A Commentary, art. 4(5) 135 s.

<sup>79</sup> Notamment la recherche en matière de *membership inference attacks*, dont plusieurs articles ont été cités auparavant.

Au vu de ce qui précède, et en admettant qu'une part majoritaire des données resteront des données personnelles au sens du droit, observons les conséquences positives que peut entraîner l'implémentation de PETs dans un système de *machine learning*, en restant dans le cadre du droit de la protection des données.

## B. Les grands principes et la protection des données dès la conception

La notion de protection des données dès la conception impose au responsable de traitement de mettre en place des mesures notamment techniques afin de garantir le respect des principes de la protection des données et ce dès la conception du traitement en question. Le lien avec les PETs semble d'emblée évident, tant leurs idées fondatrices sont intriquées. Les PETs concrétisent les préceptes de la protection des données dès la conception, bien que la recherche de technologies préservant la vie privée soit antérieure à la naissance de l'idée de protection des données dès la conception<sup>80</sup>. Partant, toute implémentation de PETs lors de la conception d'un système de *machine learning* témoigne d'une volonté de respecter l'art. 7 LPD ou 25 RGPD et, par extension, les principes de la protection des données, puisque ces derniers sont englobés dans ces dispositions. Le principe de responsabilité du RGPD, qui n'a pas d'équivalent en droit suisse, précise simplement le fait que le responsable du traitement se doit de pouvoir démontrer sa conformité, par exemple en prouvant l'implémentation de mesures techniques appropriées<sup>81</sup>. Malgré son absence de la lettre de la LPD, cette notion s'inscrit dans la continuité des principes en eux-mêmes, ainsi que de la protection des données dès la conception, et l'on peut considérer que des attentes similaires reposent sur le responsable du traitement dans le contexte suisse.

Le caractère approprié ou non des mesures attendues par les art. 7 LPD et 25 RGPD s'apprécie selon différents facteurs, dont l'approche fondée sur les risques, qui est à la base du droit de la protection des données actuel tant au niveau suisse qu'europpéen<sup>82</sup>, le type de traitement ou les coûts de mise en œuvre, notamment. Un autre facteur central est celui de l'état de la technique<sup>83</sup>, qui plus est concernant les PETs. La question de l'état de la technique est particulièrement complexe à traiter, et sera discutée

plus en détail infra (F). Le recours à la pseudonymisation est cité à titre d'exemple de mesures techniques tant par le législateur suisse qu'europpéen, ce dernier y ajoutant l'anonymisation. Comme nous l'avons dit (III B), ces techniques ont beaucoup été traitées en doctrine. En outre, elles connaissent certaines limites importantes. Cette contribution nous permet de mettre en lumière d'autres mesures techniques pouvant poursuivre le même but, avec l'espoir de le faire de manière plus réussie encore.

La confidentialité différentielle peut contribuer à respecter la limitation du traitement à sa finalité. En effet, elle peut être configurée pour ne répondre qu'à des requêtes concernant le but du traitement, sans outrepasser celui-ci, ou en implémentant plus de bruit dans les données sortant du cadre de la finalité<sup>84</sup>. Reste à savoir si l'ajout de bruit peut à lui seul être considéré comme garantissant l'absence de traitement pour d'autres finalités que celles déclarées, ou s'il doit être accompagné de mesures organisationnelles. La confidentialité différentielle peut également fortement réduire la quantité de données traitées, en accord avec le principe de minimisation des données, en augmentant le bruit en fonction si certaines données sont moins utiles pour l'objectif du système de *machine learning*<sup>85</sup>. Ceci doit aussi être nuancé en rappelant que le principe de minimisation ne se limite pas à la quantité de données collectées, mais inclut également leur pertinence et leur adéquation. Un système peut donc respecter formellement ce principe tout en soulevant des questions sur la proportionnalité de ses objectifs.

D'autre part, l'apprentissage fédéré minimise le plus possible les données traitées, celles-ci restant décentralisées, sur les dispositifs tiers et jamais en possession du responsable du traitement. Ici aussi, rappelons que la notion de minimisation va au-delà d'une simple question de volume, mais tient compte de l'adéquation des données en tant que telles. Les données fournies étant gérées par les utilisateurs eux-mêmes, le traitement peut toutefois tendre à être mécaniquement limité à sa finalité par ceux-ci.

Enfin, le chiffrement homomorphe pourrait, dans une certaine mesure, permettre de limiter le traitement à sa finalité et de minimiser le volume de données traitées en s'assurant que seules soient exposées les données utiles au but. Le chiffrement homomorphe et plus largement l'ensemble des techniques de chiffrement sont déterminantes en matière de sécurité des données, comme nous allons le voir.

<sup>80</sup> LEE A. BYGRAVE, The EU General Data Protection Regulation (GDPR), A Commentary, art. 25 573.

<sup>81</sup> QUINN (n. 13), 159 s.

<sup>82</sup> Message LPD (n. 8), 6649. En droit européen, cela ressort par exemple de l'art. 24(1) RGPD.

<sup>83</sup> Message LPD (n. 8), 6649.

<sup>84</sup> KNOCKAERT et al. (n. 35), 333.

<sup>85</sup> KNOCKAERT et al. (n. 35), 333.

### C. La sécurité des données

La sécurité des données fait partie des principes cardinaux de la protection des données. Elle est codifiée à l'art. 8 LPD, respectivement 32 RGPD. L'objectif fondamental est d'empêcher les violations de la protection des données. Les concepts de confidentialité, d'intégrité, de disponibilité et de résilience des systèmes et des services sont généralement mis en avant<sup>86</sup>. Les relations entre technique et droit sont indispensables dans ce champ, tant les notions juridiques ne trouvent de sens que par la réalisabilité technologique et l'expertise informatique. La loi n'a de portée que si le code informatique l'incorpore<sup>87</sup>.

Les PETs jouent, partant, un rôle déterminant dans ce champ aussi. L'anonymisation supprime tout risque pour les personnes concernées<sup>88</sup> en cas de fuite de données, car elles ne sont plus identifiables. La pseudonymisation, quant à elle, permet d'obfusquer une partie des données, du moins tant que la correspondance avec les données d'origine n'est pas compromise du même coup. En ce sens, on peut dire que la pseudonymisation se rapproche d'un chiffrement. On peut penser à la pseudonymisation par table hachage<sup>89</sup>, qui illustre bien la proximité entre ces deux concepts du point de vue de la sécurité des données<sup>90</sup>. La confidentialité différentielle permet de réduire la quantité de données personnelles obtenues par l'assaillant en cas d'attaque réussie, le bruit rendant inexploitable une part des données ayant fuité. La confidentialité différentielle appliquée en bout de flux de travail, pour ajouter du bruit à l'*output* d'un modèle de *machine learning*, permet également de se prémunir des attaques dites d'inférence de données (*data inference attacks*)<sup>91</sup>. L'apprentissage fédéré, ne laissant que peu, voire pas, de données entre les mains du propriétaire du modèle, rend presque inutile une attaque cherchant à extraire une base

de données chez celui-ci, puisque par définition il n'en a pas. Enfin, le chiffrement homomorphe réduit drastiquement les risques en cas de violation de données, les données étant chiffrées pendant le stockage, le transfert et les calculs<sup>92</sup>. En cas d'intrusion ou de vol de données, il n'est alors pas nécessaire d'informer les personnes concernées, car les données sont obfusquées aux yeux de l'assaillant et le risque pour celles-ci s'en trouve grandement réduit, conformément à l'art. 24 LPD et 34(3)(a) RGPD. Le chiffrement en général et notamment le chiffrement homomorphe permettent de favoriser la sécurité et la confidentialité des données. Le chiffrement fait d'ailleurs partie des mesures explicitement mentionnées par le législateur européen<sup>93</sup>.

Il apparaît clair, compte tenu des points évoqués, que les PETs sont au cœur de la sécurité des données, peut-être encore plus particulièrement que d'autres champs du droit de la protection des données.

### D. L'analyse d'impact

L'analyse d'impact, prévue respectivement aux art. 22 LPD et 35 RGPD, est une obligation du responsable d'un traitement « susceptible d'engendrer un risque élevé pour les droits et libertés des personnes concernées »<sup>94</sup>. Ce processus doit permettre de mettre en lumière certains risques d'atteinte, et les PETs peuvent être un élément central des « mesures appropriées » en réponse à ceux-ci<sup>95</sup>. Le responsable du traitement sait, sur cette base, quelles technologies en particulier implémenter et pourra prouver s'être adapté aux résultats de son analyse.

### E. L'atténuation des sanctions

Dans le cadre de la réglementation européenne, si le développeur d'un système de *machine learning* viole une disposition du RGPD et qu'une amende devait lui être imposée, l'un des facteurs permettant de déterminer le montant est l'implémentation ou non de mesures techniques appropriées au sens des art. 25 et 32 RGPD<sup>96</sup>. Ainsi, l'incorporation de solutions cherchant à préserver la protection des données dans l'architecture technique d'un modèle d'IA permet de témoigner d'une forme de « bonne volonté » qui sera récompensée (ou plutôt moins punie) en cas d'ir-

<sup>86</sup> DENNIS-KENJI KIPKER/SVEN MÜLLER, Technische und organisatorische Massnahmen (TOMs), in : Markus Kaulartz/Tom Braegelman (éd.), *Rechtshandbuch Artificial Intelligence und Machine Learning*, Munich 2020, 415 ss, 416, N 3 ss.

<sup>87</sup> Idée développée sous le fameux concept doctrinal de Code is Law, LAWRENCE LESSIG, *Code is Law – On Liberty in Cyberspace*, Harvard Magazine, 1 janvier 2000. Voir aussi CR LPD-FANTI/STAEBER, art. 8 N 3.

<sup>88</sup> Du moins qui l'étaient originellement, avant le processus d'anonymisation.

<sup>89</sup> ERDEM (n. 33), 238.

<sup>90</sup> Ces deux techniques sont d'ailleurs mentionnées ensemble à l'art. 32(1)(a) RGPD. Notons également que certains auteurs considèrent le chiffrement comme une forme particulière de la pseudonymisation, selon nous à raison. En ce sens, PC LPD-BÉGUIN, art. 7 N 6.

<sup>91</sup> YE et al. (n. 54).

<sup>92</sup> INFORMATION COMMISSIONER'S OFFICE (ICO) (n. 67).

<sup>93</sup> Art. 32(1)(a) RGPD. La LPD n'y fait en revanche pas mention directement.

<sup>94</sup> Consid. 84 RGPD.

<sup>95</sup> Consid. 84 RGPD.

<sup>96</sup> Art. 83(2)(d) RGPD ; LEE A. BYGRAVE, *The EU General Data Protection Regulation (GDPR)*, A Commentary, art. 25 578.

respect de dispositions de la loi. Cyniquement, même un développeur violant des dispositions du RGPD peut améliorer son sort du point de vue de la lourdeur de la punition en implémentant des PETs.

À la différence du RGPD, le droit suisse n'a pas ouvert la possibilité d'amendes administratives. Notons toutefois que le préposé fédéral à la protection des données et à la transparence (FPD) a le pouvoir d'ordonner des mesures administratives, parmi lesquelles figure l'obligation de mettre en place des mesures techniques pour préserver la protection des données des personnes concernées par le traitement, qui pourrait donc être l'implémentation de PETs<sup>97</sup>. En revanche, quant au volet pénal, ne pas respecter les exigences minimales en matière de sécurité des données de l'art. 8 al. 3 LPD de manière intentionnelle peut entraîner, sur plainte, une amende pouvant s'élever jusqu'à 250 000 francs<sup>98</sup>. Ces exigences minimales ont été détaillées par le Conseil fédéral dans l'ordonnance relative à la loi sur la protection des données<sup>99</sup> et une approche législative fondée sur le risque a également été adoptée, à l'instar de la protection des données dès la conception<sup>100</sup>. Partant, et compte tenu de l'impossibilité de prévoir tous les cas sans tomber dans l'excès réglementaire, ce sont des lignes directrices qui ont été prévues, afin d'aider à déterminer quelles mesures sont appropriées au cas d'espèce<sup>101</sup>.

Mettre en place des PETs dans la conception d'un système de *machine learning* peut ainsi réduire les risques de se voir opposer une mesure administrative ou d'être condamné à une amende, ou encore (pour le cas du RGPD) peut réduire la magnitude de celle-ci.

## F. La question de l'état de la technique et de la proportionnalité

Tous les points traités jusqu'ici sont, notamment par le biais de la notion de proportionnalité, dépendants de l'état de la technique et des coûts supposés par l'implémentation. En adoptant une posture très conservatrice, on pourrait rapidement dériver vers des exigences démesurées envers les responsables de traitement ou les développeurs de systèmes de *machine learning*, en attendant d'eux une implémentation de dizaines de techniques ultramodernes encore peu testées et très coûteuses à mettre en place<sup>102</sup>, sans égards au produit final ni à la réalisabilité du projet.

Dans les faits, une telle interprétation du droit de la protection des données mènerait à l'impossibilité de développer et de déployer des systèmes d'IA, ce qui n'est probablement pas la volonté du législateur européen ou suisse, même lorsqu'ils veillent à garantir un cadre juridique favorable à l'innovation<sup>103</sup>. La subtile nuance permettant de manœuvrer ce droit réside dans la balance entre protection la plus forte possible et mesures exigibles, balance fondée sur la proportionnalité. Cet équilibre est difficile à trouver et, pire encore, est en évolution constante. Il est probablement disproportionné d'attendre aujourd'hui des développeurs la mise en place d'un apprentissage fédéré couplé de chiffrement homomorphe dans un système multiplateformes et destiné à un nombre très large d'utilisateurs, mais ceci pourrait devenir raisonnable dans les prochaines années. Ce défi est de la responsabilité du juge, d'une part, et de l'exécutif, d'autre part, ainsi que peut-être à plus long terme d'évolutions législatives judicieuses.

Notons que, prenant le contre-pied de ce que nous venons d'admettre, il est possible d'argumenter que ce n'est pas le droit de la protection des données qui est trop ferme et peu permissif, mais plutôt le *machine learning* en tant que tel qui est trop contraire à la loi, de par sa gourmandise en données, ses dangers liés au profilage de masse et son caractère général et transversal, et que, plutôt que de céder à la pression du progrès, il pourrait être souhaitable de freiner la machine. À nouveau, trancher ce débat est éminemment politique, voire philosophique, et y donner une réponse définitive dépasse tant l'objet de la présente contribution que les limites de nos compétences.

Remarquons ainsi qu'il est difficile de donner une réponse *de lege lata*, car une large part des normes prévues par le droit de la protection des données sont formulées de manière larges et soumises à interprétation, cette dernière pouvant fortement varier en fonction de l'opinion sur ces débats<sup>104</sup> de la personne qui les interprète.

<sup>97</sup> Art. 51 al. 3 let. b LPD.

<sup>98</sup> Art. 61 let. c. LPD.

<sup>99</sup> Ordonnance sur la protection des données du 31 août 2022 (OPDo ; RS 235.11).

<sup>100</sup> CR LPD-FANTI/STAEGER, art. 8 N 94.

<sup>101</sup> CR LPD-FANTI/STAEGER, art. 8 N 94 ss.

<sup>102</sup> PC LPD-BÉGUIN, art. 8 N 15 *in fine*.

<sup>103</sup> PC LPD-BÉGUIN, art. 8 N 15 *in fine*. Notons que la notion de « bac à sable » réglementaire du RIA permettrait l'utilisation des données à caractère personnelle collectées à d'autres fins pour le développement de certains systèmes d'IA considérés d'intérêt public, sous le contrôle de l'autorité compétente et uniquement dans les conditions déterminées par l'art. 59 RIA. Ces conditions prennent en considération les aspects fondamentaux du RGDP (consid. 140 RIA).

<sup>104</sup> Esquissés aux deux paragraphes précédents.

## V. Conclusion

Afin de continuer de profiter des avantages offerts par l'IA, sans pour autant mettre en péril la personnalité des personnes dont les données sont utilisées pour entraîner ces systèmes, ni en ignorant les risques à l'échelle de la société dans son ensemble, l'une des solutions potentielles est la mise en place de mesures techniques imposant le respect des dispositions du droit de la protection des données. La pseudonymisation et l'anonymisation, la confidentialité différentielle, l'apprentissage fédéré ou le chiffrement homomorphe, mais également les autres PETs qui n'ont pas été discutés ici, sont autant de pistes de réflexion pour la recherche, avec comme finalité une IA respectueuse de nos données. Peut-être encore un peu vertes pour les voir être implémentées dans tous les systèmes de *machine learning* dès aujourd'hui, on pourrait voir mûrir les PETs au cours des prochaines années, en fonction des avancées de la recherche, mais aussi de l'évolution des conceptions juridiques en matière de protection des données.

De tels changements ouvriront irrémédiablement de nouvelles questions, *de lege ferenda*, auxquelles les experts en droit et en informatique devront tenter de répondre, main dans la main. Les domaines d'expertise ne peuvent plus travailler en silos fermés en matière d'enjeux numériques, car ces derniers sont fondamentalement pluridisciplinaires et transversaux. Dans cette perspective, il est essentiel de renforcer les collaborations entre chercheurs, praticiens et régulateurs afin d'anticiper les défis émergents et d'assurer un développement technologique éthique et conforme aux droits fondamentaux. Ce travail collectif, loin de freiner l'innovation, peut devenir un levier pour concevoir des systèmes d'intelligence artificielle à la fois performants, responsables et dignes de confiance, aptes à répondre aux attentes d'une société numérique en constante évolution.