Pour un développement des IAs respectueux de la vie privée dès la conception

Maël Pégny

POUR UN DEVELOPPEMENT DES IAS RESPECTUEUX DE LA VIE PRIVEE DES LA CONCEPTION

Maël Pégny, Projet OLKi

POUR UN DEVELOPPEMENT DES IAS RESPECTUEUX DE LA VIE PRIVEE DES LA CONCEPTION .. 1
MAËL PEGNY, PROJET OLKI 1
PRINCIPES ETHIQUES POUR LES DEVELOPPEURS – LA CHARTE 2
PRINCIPES ETHIQUES POUR LES DEVELOPPEURS – LE COMMENTAIRE 4
Introduction : objet, thème, public, style 4
Fonction de la charte 6
I. CONCEPTION DE L'OBJET : POUR UN DEVELOPPEMENT ETHIQUE TOUT AU LONG DU CYCLE DE VIE DES MODELES 8
Délimitation de l'objet et état actuel du droit : le problème de la finalité du traitement .. 12
II A. FAILLES DE SECURITE DES MODELES ET RESPECT DE LA VIE PRIVEE : UN ETAT DE L'ART RAISONNE AUX FINS DE DISCUSSION ETHIQUE 14
L'inversion des modèles 14
Les attaques par inférence d'appartenance (membership inference) 14
Les excès du pouvoir prédictif 15
La lutte contre la suroptimisation 16
Le pouvoir d'inférer des données personnelles à partir de données publiques 16
II B SOLUTIONS POSSIBLES AUX PROBLEMES DE RESPECT DE LA VIE PRIVEE 17
Pour un entraînement des modèles respectueux de la vie privée 17
1. L'exclusion des données personnelles des bases d'entraînement : de la définition des données problématiques à l'exclusion des données personnelles de d'apprentissage 18
2. L'approche de sécurité post hoc 25
2.1 Une autre approche : la certification par la vérification 25
2.2. L'acceptation de la restriction d'accès pour des cas exceptionnels 28
3. Le désentraînement des modèles 29
4. La confidentialité différentielle 29
III. CONSEQUENCES PHILOSOPHIQUES ET JURIDIQUES DE L'ETAT DE L'ART 31
Le brouillage de la distinction entre données et logiciel 31
5. L'état de l'art juridique sur le concept de donnée personnelle 32
La remise en cause de la notion de donnée personnelle 34
Le droit à l'inférence raisonnable 36
Conclusion 37

1. **Dans le cadre de recherches scientifiques, déclarer les finalités de l'usage des données et l'extension de la collecte nécessaire à ces finalités, documenter et justifier tout écart à cette déclaration initiale par une découverte faite dans le cours de la recherche et l'intérêt scientifique d'une modification des hypothèses. Discuter explicitement les impacts possibles sur la vie privée de ces changements de finalité et de collecte.**

2. **Afin de prévenir l'apparition d'un pouvoir de prédiction trop fin présentant une menace pour la vie privée, tester les performances finales du logiciel en s'interrogeant sur la nécessité des résultats par rapport aux objectifs d'apprentissage et aux finalités du traitement. Prendre toutes les mesures possibles pour limiter le pouvoir prédictif du modèle à ce qui est strictement nécessaire pour ces objectifs et finalités.**

3. **Dans les arbitrages sur la terminaison du processus d'apprentissage, et le compromis à trouver entre risque de suroptimisation et perte de performances, prendre en compte les risques posés par la suroptimisation pour la vie privée.**

4. **Entraîner son modèle sans avoir recours à des données personnelles. Si ce principe est inapplicable ou sans pertinence, examinez les options offertes par les principes 5 ou 6 selon le contexte technique et éthique.**

5. **Entraîner son modèle sans faire usage de données personnelles dont la diffusion pourrait porter atteinte aux droits des personnes.**

6. **Entraîner son modèle en faisant exclusivement emploi de données ayant fait l'objet d'un geste explicite de publication. Mettre en place toutes les mesures possibles, à la fois automatiques et manuelles, pour permettre la mise à jour des données, afin de prendre en compte la correction de données erronées, le retrait de publication et l'exercice des droits de rectification, d'effacement et d'opposition.**

7. **Si le recours à des données personnelles est inévitable, déclarer les raisons justifiant ce recours, ainsi que toutes les mesures prises pour lutter contre la rétro-ingénierie des données, et prendre position sur leur complétude par rapport aux boites à outils et aux méthodes d'attaque existantes.**

8. **Diffuser en licence libre tous les outils de sécurisation contre la rétro-ingénierie des données.**

9. **Si cela n'entraîne pas de faille de sécurité intolérable, mettre le modèle à disposition de tous afin de permettre la vérification publique des propriétés de sécurité. Détailler les raisons de la décision positive ou négative, les mesures de sécurité prises contre les failles créées par cette mise à disposition du public, et prendre position sur leur complétude par rapport à l'état de l'art.**

10. **Lorsque les mesures de restrictions de la collecte et de lutte contre la rétro-ingénierie ne sont pas applicables, et que la gravité de l'enjeu surpasse les enjeux de vie privée, autoriser un modèle encodant des données privées en restreignant strictement l'accès à ce modèle et son emploi pour l'usage ayant justifié l'exception. La discussion de l'exception faites au respect de la vie privée et à la publicité de la connaissance scientifique devra prendre en compte les propriétés singulières des modèles de ML, comme la capacité à apprendre en temps réel sur une grande masse de données, l'opacité du fonctionnement et de son évolution, et la qualité de leurs prédictions comparée à d'autres modèles.**

Principes éthiques pour les développeurs – le commentaire

Introduction : objet, thème, public, style

Il existe de (trop) nombreuses chartes consacrées à l'éthique de l'IA, mais la plupart d'entre elles sont centrées sur l'éthique de l'usage de ces IAs considérées comme des produits finis et les nombreux impacts de cet usage sur nos sociétés. La présente charte est consacrée à l'éthique du développement, et à l'impact que ces pratiques du développement peuvent avoir sur les autres questions. En plus de se spécialiser sur un objet, elle se spécialise sur un thème, à savoir les problèmes de respect de la vie privée posés non seulement par la collecte massive de données nécessaire au développement, mais aussi par l'enregistrement, et la possible récupération de données à partir des modèles. Cet enjeu nous semble d'autant plus important qu'il est propre aux modèles d'apprentissage automatique, et il constitue par conséquent une question nouvelle suscitée par les évolutions technologiques les plus récentes.

À ces choix d'objet et de thème correspond nécessairement un choix de public. Cette charte s'adresse en premier lieu aux développeurs et aux institutions productrices d'IAs, non seulement parce que c'est leur conscience morale et politique qui est ici en jeu, mais parce que notre travail suppose de rentrer dans des considérations techniques qui sont tenues à distance, de manière bien compréhensible, par la plupart des chartes généralistes. Nous ne souhaitons pas en effet nous restreindre à la définition de valeurs à respecter, mais voulons initier le travail d'opérationnalisation de ces valeurs dans la pratique même du développement, qui seule permettra de ne pas demeurer au stade de pieuses déclarations d'intentions, et nourrira la réflexion éthique de la richesse des difficultés pratiques. Nous rejoignons aussi l'approche défendue par l'*Association of Internet Researchers* dans la troisième version de sa charte[1], selon laquelle il faut à tout prix rompre avec une vision de l'éthique comme quelque chose qui se fait avant la recherche, l'étouffe et la limite, et consiste essentiellement à cocher des cases. La réflexion éthique doit être vue comme quelque chose de dynamique qui nourrit une meilleure conception de la recherche et mène à quelque chose de scientifiquement plus mûr. Ceci est particulièrement vrai pour le respect de la vie privée, qui peut facilement être réduite à un corset réglementaire restreignant la collecte et le traitement de données, un ensemble de mesures de sécurité et de tâches bureaucratiques. L'existence de contraintes réglementaires n'est en rien un mal lorsqu'il s'agit de protéger des droits fondamentaux, mais nous verrons que l'éthique du développement pose des questions bien plus fines qui ne peuvent se penser purement en termes de contraintes externes.

Cette charte s'adresse en outre à un public particulier de développeurs, attaché aux valeurs de publicité de la science et du logiciel, et d'un développement éthique respectueux des utilisateurs et soucieux de l'impact du numérique sur la société. Ce public regroupe des communautés relativement dispersées, comme celles des partisans du logiciel libre, des communs numériques, de la reproductibilité de la recherche et de la conception logicielle éthique. Il nous semble que les évolutions techniques posent des enjeux particuliers à ce public. La publicité du logiciel doit maintenant affronter les défis inédits posés au respect de la vie privée par la publication des modèles d'apprentissage automatique. L'enjeu du respect de la vie privée n'est pas nouveau, puisqu'il n'existe pas de licence libre pour un document contenant des informations personnelles. Lorsqu'on défend, selon une analogie si fréquente dans la littérature, que les logiciels sont comme des recettes de cuisine qui appartiennent à tous, on ne songe pas aux problèmes singuliers posés par des modèles qui ne contiennent pas que des algorithmes, mais aussi un encodage de données parfois personnelles, et parfois sensibles. Le problème est donc qu'ici c'est le logiciel lui-même, l'objet central de la culture libriste, qui devient potentiellement un document contenant des informations personnelles, c'est-à-dire un artefact que cette culture même a considéré comme extérieur à son

[1] Aline Shakti Franzke et al., « Internet Research: Ethical Guidelines 3.0 » (Association of Internet Researchers, 2020), https://aoir.org/reports/ethics3.pdf.

objet. L'une des fins de cette charte est donc de trouver une voie de conciliation entre l'idéal de publicité du logiciel et celui de respect de la vie privée à l'égard de ces modèles. Une telle conciliation est d'autant plus urgente que la publicité n'est pas nécessaire seulement pour un idéal de diffusion des connaissances informatiques comme bien commun, mais aussi comme moyen de contrôle du respect des principes éthiques. La publicité permet la transparence et la vérification de la loyauté des déclarations faites sur les fonctionnalités : elle permet de vérifier que le logiciel a bien les propriétés qu'il est censé avoir, et fait bien ce qu'il est censé faire. Mais comment opérer quand cette activité de vérification publique peut mettre en danger la valeur même que l'on souhaite vérifier, à savoir le respect de la vie privée ? Comment développer une culture de la sécurité des modèles qui soit compatible avec la culture libriste de la publicité ? Les modèles d'apprentissage automatique sont voués à devenir un point de tension entre le désir de respect de la vie privée et les idéaux de diffusion libre du logiciel et de reproductibilité de la recherche, et c'est cette tension que cette charte veut contribuer à résoudre.

D'un point de vue encore plus général, les enjeux liés aux respects de la vie privée et à l'IA touchent au cœur des discussions sur les communs numériques. Les communs de la connaissance sont parfois présentés comme des biens publics parfaits, à l'opposé des biens communs physiques qui supposent toujours une part de rivalité et d'excluabilité. Ils ouvriraient la possibilité de « reproduction indéfinie de biens rivaux à l'identique et sans perte d'information et pour un coût quasi-nul [2] ». Les problèmes liés à la vie privée viennent cependant apporter une limitation à cette définition, car les informations personnelles ne peuvent être partagées sans un effet décisif qui change leur statut. Ce qui fait la valeur qualitative d'une information privée, c'est précisément la limitation de sa diffusion : sa diffusion aux quatre vents constitue donc une dégradation et non une ouverture à un épanouissement collectif. En revanche, la lutte contre l'appropriation des données personnelles est présentée par le même manifeste comme une logique d'enclosure qui dépossède les utilisateurs de leurs propres données, ce qui montre que la culture de préservation des données personnelles ne constitue rien d'étranger aux communs numériques. Mais le problème est que l'enclosure est habituellement définie comme une appropriation privée forcée d'un commun, soit d'un bien collectif, et non justement d'un bien par définition individuel comme les informations personnelles. Qu'elle soit ou non conçue comme une forme de propriété privée, la vie privée se doit de rester privée. Il faut donc pouvoir développer un cadre philosophique permettant la coexistence de communs de la connaissance avec le respect de la propriété privée des données personnelles. Comme dans le cas particulier de la culture libriste, les IAs vont poser problème car elles sont à la fois des objets de connaissance vouées à devenir des communs numériques, et de possibles dépositaires d'informations personnelles voire privées : elles constituent donc un véritable nœud gordien de la culture des communs de la connaissance numérique.

Nous souhaitons cependant que cette charte puisse être intéressante à lire pour des développeurs n'adhérant pas à la culture libriste, ou n'étant pas en mesure de la pratiquer dans leur vie professionnelle. C'est la raison pour laquelle nous accordons la plus grande importance à la formulation des problèmes moraux fondamentaux, notamment ceux impliquant un arbitrage entre des valeurs incommensurables. Quelle que soit la position prise sur ces arbitrages notoirement difficiles à rationaliser, tous peuvent gagner à une claire formulation des questions, et une nette délimitation de l'espace des possibles.

[2] C'est par exemple l'approche défendue par le Manifeste du collectif SavoirsCom1, que nous citons ici : « Le manifeste de SavoirsCom1 | SavoirsCom1 », consulté le 6 septembre 2020, https://www.savoirscom1.info/manifeste-savoirscom1/.

Fonction de la charte

Nombre de lecteurs peuvent être sceptiques à l'égard de l'exercice même de la rédaction d'une charte, et nous souhaitons faire face à ce scepticisme[3]. Les chartes sont intrinsèquement limitées par leur absence de force contraignante. Face à des pratiques bien ancrées et à l'opposition de pouvoirs institutionnels et d'intérêts économiques immenses, les déclarations éthiques peuvent sembler vaines, délibérément dépourvues de prises de positions fortes[4], voire servir de cache-sexe éthique à des institutions luttant activement contre les évolutions éthiques que nous voulons stimuler. L'éthique peut ainsi être instrumentalisée par les institutions, et notamment les géants du numérique, à la fois comme instrument -pour orienter, limiter et façonner le débat public sur ces questions de la manière la plus favorable à leurs intérêts, y compris lorsque ceux-ci peuvent être détestables d'un point de vue éthique. Il ne s'agit malheureusement pas que d'une inquiétude de principe. Lorsque l'on voit que Google publie une charte plaçant le respect de la vie privée dès la conception (*privacy by design*) au cœur de ses objectifs[5], ou que l'Académie chinoise de l'Intelligence Artificielle se fend d'une charte pleine de bonnes intentions éthiques[6], on ne sait plus si l'on doit rire ou pleurer. Au mieux impuissantes, au pire néfastes par leur instrumentalisation comme écran de fumée, les chartes seraient donc un exercice scolastique dont on ferait mieux de se dispenser, pour se consacrer à des objectifs politiques plus solides.

Nous tenons à rappeler que la rédaction d'une charte n'est en aucun cas exclusive d'une évolution contraignante du droit, de la création d'autorités de contrôle, ni d'aucune autre forme d'action politique[7]. Il ne s'agit en aucun cas pour nous de fournir aux institutions productrices d'IAs un chèque en blanc éthique, qu'il leur suffirait de signer pour se déclarer soldées de tout compte. Mais avant d'aborder les questions de stratégie politique et institutionnelle complexes que suppose une transformation profonde des pratiques, encore faut-il savoir après quoi l'on court. Or, il nous semble manifeste que la conscientisation de la communauté des développeurs est à cet égard très loin de ce qu'elle devrait être même au simple niveau de la position des problèmes, sans même parler de l'implémentation technique d'une solution dans la pratique du développement. Nombre de développeurs, dans l'industrie comme dans la recherche publique, collectent toutes les données sur lesquelles ils peuvent mettre la main sans même avoir une conscience claire de l'état actuel du droit, des enjeux éthiques soulevés par leur activité, des questions de sécurité posés par la possession de ces données et leur enregistrement dans des modèles. La première fonction de cette charte est donc de provoquer une prise de conscience de ces enjeux, d'inciter à la prise de position, et d'initier dans la communauté le nécessaire débat sur les valeurs que nous voulons implémenter dans notre pratique, les méthodes permettant leur opérationnalisation et les limites techniques et

---

[3] Pour un exemple d'un tel scepticisme, voir Thilo Haggendorf, « The Ethics of AI Ethics », *Minds and Machines* 30(1) : 99-120, 2020, Url : https://arxiv.org/ftp/arxiv/papers/1903/1903.03425.pdf

[4] On a pu ainsi regretter l'absence de lignes rouges contre les systèmes d'armements léthaux autonomes ou les systèmes de scoring social dans Groupe d'Experts Indépendents de Haut Niveau sur l'IA, « Lignes directrices en matière d'éthique pour une IA digne de confiance » (Commission Européenne, 2019). Les objections, et leurs références, sont mentionnés dans le document lui-même, dans un geste de franchise très rare dans ce type de rapports officiels.

[5] « Our Principles », Google AI, consulté le 6 septembre 2020, https://ai.google/principles/.

[6] « Beijing AI Principles », consulté le 6 septembre 2020, https://www.baai.ac.cn/news/beijing-ai-principles-en.html. Il a été remarqué que, malgré sa considérable hypocrisie, ce document n'est pas dénué de sens politique dans le contexte politique chinois. De par sa grande similarité avec nombres de chartes publiées en Occident, et son insistance sur l'harmonie et la coopération internationales, il montre une volonté de dialogue sur les questions d'éthique de l'IA de la part du gouvernement chinois qui n'était pas une évidence *a priori,* même si ceci ne change évidemment rien à l'usage massif de l'IA à des fins de surveillance et de répression. Voir Will Knight, « Why Does Beijing Suddenly Care about AI Ethics? », *MIT Technology Review*, 31 mai 2019, https://www.technologyreview.com/2019/05/31/135129/why-does-china-suddenly-care-about-ai-ethics-and-privacy/.

[7] La même position est adoptée par le Groupe d'Experts Indépendants de Haut Niveau sur l'IA, GENH IA, dans ses *Lignes directrices pour une IA digne de confiance* : « Les présentes lignes directrices ne visent ni à remplacer toute forme actuelle ou future d'élaboration de politiques ou de réglementations ni à en décourager l'introduction. »

éthiques de cette même opérationnalisation. Ce débat doit avoir lieu avant que nous affrontions les problèmes politiques considérables que posent une transformation des pratiques, si nous voulons éviter d'être victimes d'une définition floue de nos objectifs politiques, d'une faible conscientisation et mobilisation des premiers concernés et d'une maigre expérience opérationnelle. Il doit être nourri au plus vite par les retours d'expérience des développeurs qui tâcheront d'implémenter les principes dans leur pratique, afin de permettre l'accumulation de connaissances qui donnera une véritable quantité de mouvement à la démarche.— Cette charte n'est donc pas une fin, mais un commencement.

L'une des conversations que nous essayons de lancer est une conversation sur les limites de l'opérationnalisation des concepts éthiques. Pour cette raison, nous n'hésiterons pas à faire des propositions de principes parfois très éloignés de la pratique actuelle. Tout d'abord, un principe qui ne serait applicable que dans 2% des cas n'est pas forcément indigne d'être discuté, si jamais ces 2% pouvaient contenir des cas d'usage très sensibles. En outre, la discussion sur l'opérationnalisation des principes n'est pas intéressante que par ses résultats, mais aussi par les raisons qui justifient l'impossibilité d'opérationnaliser telle ou telle approche. Il serait donc malheureux d'exclure d'emblée des principes de toute considération parce qu'ils semblent durs à opérationnaliser, alors que ce sont précisément ces difficultés que nous voulons faire ressortir dans le débat.

On peut aussi dire que, loin d'être exclusive d'une évolution du droit, la fonction d'une charte est de pousser la réflexion normative au-delà des objets déjà traités par le droit, et d'inciter ainsi juristes et législateurs à prendre conscience des limites de l'état de l'art juridique et à envisager la conquête de nouveaux territoires. Cette conception de la fonction de la charte définit immédiatement une restriction de notre objet, qui doit être centré sur les aspects qui ne sont pas déjà traités par le droit positif. Le droit positif, notamment européen, contient nombre de principes et règles puissants qui, même s'ils peuvent être entachés de nombreuses exceptions et problèmes d'interprétation, devraient déjà se voir donner une chance d'être appliqués rigoureusement. Le consentement et re-consentement, le principe de communication claire et explicite des modalités du traitement, la sécurisation des données, la minimisation de la collecte, le droit à l'explication, le principe d'exactitude des données personnelles, qui contient déjà une obligation intrinsèque de mise à jour car les données doivent être « exactes et, si nécessaire, tenues à jour… », le principe de conservation limitée des données durant le temps minimal nécessaire au traitement, etc. font déjà partie du droit européen. Nous nous concentrerons donc en priorité sur les problèmes qui sont à notre sens incomplètement pris en compte par l'état de l'art juridique.

Lorsque nous évoquerons ici le droit positif, nous nous consacrerons exclusivement au droit européen tel qu'il est aujourd'hui exprimé dans le Règlement Général de Protection des Données (RGPD), et les différents commentaires et éléments de jurisprudence sur cette législation. Il y a trois raisons à cette restriction. La première est que le RGPD est souvent considéré comme l'étalon juridique mondial en ce qui concerne la protection des données personnelles. La seconde est que le RGPD a une influence qui excède largement les frontières de l'UE. Soit qu'ils soient impressionnés par cette approche, soit qu'ils désirent simplement pouvoir commercer aisément avec l'Union Européenne, les législateurs de très nombreux pays font entrer dans leur droit une vision similaire[8]. La troisième est que la prise en considération des problématiques de droit comparé

[8] Pour une vaste étude du droit comparé sur la protection des données, voir Graham Greenleaf, Global Data Privacy Laws 2017. 120 pays de l'ensemble des aires culturelles ont maintenant une législation protégeant les données personnelles, et l'approche européenne, notamment par la Convention 108, a gagné des partisans ces dernières années. Pour un site mis à jour sur l'état de la législation de protection des données, voir *DLA PIPER, Data Protection Laws in the World :* https://www.dlapiperdataprotection.com/index.html?t=definitions%26c=ZA%26c2 On peut remarquer que le patron d'un géant américain de l'industrie comme Mark Zuckerberg a pu récemment appeler les États-Unis à adopter une législation plus similaire du RGPD. Mark Zuckerberg, « Opinion | Mark Zuckerberg: The Internet Needs New Rules. Let's Start in These Four Areas. », *Washington Post*, 30 mars 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-

aboutirait à donner une trop grande place à des considérations très techniques. Nous ne sommes tout d'abord pas compétents pour mener une telle étude juridique, et nous souhaitons avant tout nous consacrer à la discussion de concepts éthiques fondamentaux, et l'approche forte du droit européen contient largement assez de matière pour mener une telle discussion. Cela ne signifie évidemment pas que nous considérions comme dépourvue de pertinence la comparaison avec d'autres droits, d'autres situations concrètes et d'autres cultures. Au contraire, cette comparaison est une part essentielle du débat que nous essayons ici de lancer.

Nous procéderons tout d'abord par un approfondissement de la discussion de la conception de notre objet (section I). Nous poursuivrons par un état de l'art raisonné des failles de sécurité affectant la protection des données personnelles en *Machine Learning* (ML), en nous concentrant sur le problème des attaques par inversion (section II.A), avant de lister les principales contremesures existantes (section II.B). Nous discuterons ensuite les problèmes éthiques posés par ces nouveaux enjeux de sécurité et leurs solutions techniques, en nous appuyant sur l'état du droit (section III).

I. Conception de l'objet : pour un développement éthique tout au long du cycle de vie des modèles

Notre approche est centrée sur l'éthique du développement, une éthique pour, dans et par la conception[9], et doit donc être contrastée avec une approche centrée sur l'usage des techniques. Pour exemple d'une telle approche, on peut prendre la *Déclaration de Montréal*[10]*,* qui défend essentiellement l'existence d'espace dénué de surveillance et la possibilité de se déconnecter. Lorsque le contrôle de ses données par l'utilisateur est évoqué, les modalités techniques de ce contrôle, et les nouveaux défis que posent à cet égard les attaques par inversion, ne sont pas discutées.

Un autre angle d'attaque du respect de la vie privée dans l'IA est évidemment d'élargir la question de l'usage pour inclure tous les dangers politiques représentés par la collection ubiquitaire des données, la centralisation sans précédent de l'information provoquée par le Big Data et les nouvelles possibilités d'exploitation, parfois délétères, créés par le ML dernière vague. À titre de (contre)-exemple, dans sa présentation *Big Data & Sustainable Development Goals,* les Nations Unies identifient essentiellement des usages de l'IA qu'elles voient comme compatibles avec leurs objectifs de développement durable, avec une présentation systématiquement positive de certaines technologies et aucune évocation des dangers politiques ou de vie privée. L'accumulation des applications citées mènerait pourtant à une gigantesque datafication des activités et centralisation des données[11]. —Les solutions techniques proposées comprennent ainsi des collectes et centralisations des données financières, agricoles, de l'expression des sentiments dans les médias et de transport, pour n'en citer que quelques-unes, si bien qu'on ne peut guère que s'inquiéter que le paradis du développement durable soit aussi celui de la surveillance généralisée.

On voit donc par contraste l'intérêt d'une approche qui ne se contente pas de citer des usages et leur éventuel compatibilité avec une finalité politique donnée, mais comprenne une analyse

areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.. Que cet appel soit sincère ou non, il ouvre néanmoins une possibilité bien réelle d'évolution en ce sens.

[9] On peut trouver cette formulation dans la version courte des *EU Guidelines for Trustworthy AI,* même s'il nous semble que le document dépasse *de facto* le cadre d'une telle approche.

[10]« La Déclaration de Montréal en IA responsable », consulté le 6 septembre 2020, https://www.declarationmontreal-iaresponsable.com/la-declaration.

[11] Par contraste, le rapport de la Commission éthique du Ministère allemand des transports précise que la centralisation des données des véhicules autonomes et connectées, et la mise en place d'une infrastructure de contrôle centralisé, ne peut se faire que s'il est possible de mettre en place des mesures empêchant toute surveillance généralisée.« Report by the Ethics Commission on Automated and Connected Driving » (Federal Ministry of Transport and Digital Infrastructure, juin 2017), https://www.bmvi.de/blaetterkatalog/index.html?catalog=382378#page_1.

politique globale selon plusieurs dimensions. Cependant, un traitement exhaustif de ces enjeux politiques titanesques dépasserait de loin le cadre de cette charte. Nous nous concentrons ici sur la discussion de l'approche de respect de la vie privée dans le développement même, et sa capacité à faire face aux nouveaux défis posés par l'encodage de données au sein des modèles. Par exemple, l'automatisation de la reconnaissance faciale et de la reconnaissance faciale en direct (*live facial recognition*), une technologie toute droit sortie des derniers progrès du ML en reconnaissance de formes, ainsi que la conception d'immenses bases de données contenant parfois des milliers de personnes jamais condamnées[12] posent évidemment des problèmes politiques immenses, qui incluent mais ne se réduisent pas au respect de la vie privée. Mais ces problèmes sont essentiellement liés à l'usage des technologies plus qu'à leur conception, et sont donc hors de la portée de ce document.

Sous un autre angle, les questions de respect de la vie privée sont souvent évoquées en même temps que les questions d'accès aux données et de qualité et d'intégrité des données. L'intégrité des données, notamment les attaques par pollution délibérée des données d'entraînement, et les questions d'accès, par exemple les mécanismes de contrôle permettant de savoir qui accède quand aux données et dans quel but[13], sont évidemment centrales pour les questions de respect de la vie privée. La préservation de la vie privée ne provient pas que de la restriction de la diffusion de données personnelles correctes, elle peut aussi provenir de la diffusion d'informations incorrectes, en particulier, mais pas uniquement, d'informations portant atteinte à la réputation des personnes. Néanmoins, nous ne souhaitons pas nous engager dans une discussion générale des mesures propres à garantir la qualité scientifique du développement en IA. La qualité des données est évidemment une question scientifique centrale de l'IA, et nous ne pouvons l'évoquer que selon un angle particulier. Nous ne souhaitons pas non plus discuter toutes les mesures de sécurité préservant l'accès aux données et leur intégrité, ce qui nous embarquerait dans un débat technique spécialisé. L'approche de respect de la vie privée dès la conception comprend bien sûr toutes ces mesures de sécurité, mais si une telle approche ne constitue pas une sous-branche de la sécurité, c'est bien qu'elle va au-delà des simples enjeux de sécurité *stricto sensu,* et c'est cet au-delà que nous voulons explorer.

Même si nous ne souhaitons pas écrire un document politique général sur les enjeux de vie privée, il nous faut cependant préciser notre conception du rôle de l'approche éthique dès la conception dans le combat politique pour le respect de la vie et les libertés politiques dans un monde soumis à la datafication ubiquitaire. L'approche éthique dès la conception en général, et l'approche de respect de la vie privée dès la conception en particulier, ne doivent en aucun cas servir d'excuse pour ne pas réfléchir aux enjeux politiques qui les dépassent de beaucoup. L'emploi de telles approches ne constitue pas plus un solde de tout compte politique que la signature d'une charte. L'une des limites les plus fondamentales des approches éthiques dès la conception a été bien mis en lumière par le rapport de la commission Éthique du Ministère fédéral allemand des transports sur les véhicules autonomes et connectés. Selon ce rapport, les particularités des cas éthiques ou juridiques ne peuvent être complètement comprises par une approche *ex ante* générale et abstraite, et donc être inclus dans la programmation[14]. Il y a une irréductibilité de la casuistique éthique et juridique au raisonnement planificateur du programmeur, irréductibilité enracinée dans des limitations épistémiques fondamentales de la programmation, d'où la nécessité d'une autorité publique pour tirer les leçons des accidents. Cependant, on peut envisager que l'approche éthique du développement soit plus vaste que la simple approche d'éthique dès la conception, car elle peut inclure le *feedback* de la casuistique morale dans les opérations de maintenance, de mise à jour et de modification. Une telle approche, dans le cas du respect de la vie privée comme des autres valeurs,

[12] Voir le rapport sur les technologies de reconnaissance faciale du comité la *House of Commons* britannique : Jennifer Brown et al., « Facial recognition and the biometrics strategy » (House of Commons, 30 avril 2019).

[13] Voir les *Ligne directrices* pour plus d'informations à ce sujet.

[14] *Ibid,* point 8, p.7

a le mérite d'évacuer le fantasme technocentriste d'une délégation complète des questions éthiques aux développeuses et développeurs, tout à coup bombardés à leur insu Grands Planificateurs de la Vie Morale.

Ce fantasme technocentriste est précisément celui sur lequel pourraient miser les institutions pour se voir décerner un blanc-seing éthique, et instrumentaliser ce blanc-seing pour esquiver leurs responsabilités ou dissimulées des pratiques douteuses intervenant en aval du développement, sur l'air de « notre modèle ne peut poser de problèmes éthiques puisqu'il est éthique dès la conception. » Ce fantasme de la résolution de tout problème éthique en amont de l'usage doit être combattu, et se voir substituer une approche de développement éthique parcourant tout le cycle de vie du logiciel, et capitalisant sur les leçons de l'usage pour améliorer la conception. Cette approche du développement éthique justifie une extension de la portée de la réflexion, qui ne doit pas se limiter au seul système technique *stricto sensu,* mais doit prendre en compte le système sociotechnique entourant l'IA. Ainsi, le GENH IA tente une première ébauche de définition de ces systèmes sociotechniques[15] en affirmant que « [c]es systèmes se composent d'êtres humains, d'acteurs étatiques, d'entreprises, d'infrastructures, de logiciels, de protocoles, de normes, de gouvernance, de législations existantes, de mécanismes de contrôle, de structures d'incitation, de procédures d'audit, de meilleures pratiques, de documentation, et d'autres éléments. » Il serait ici bon de s'inspirer de la culture de la sécurité qui comprend une composante humaine et une composante technique en constante interaction. En prenant en compte le vaste paysage d'interactions créé par l'inscription du logiciel dans un système sociotechnique, on améliore la capacité du développeur à comprendre le cycle de vie de l'usage qui est fait de son système, et à mieux en comprendre les défis à la fois *ex ante* et *post hoc.* C'est particulièrement vrai quand on discute la possibilité d'inclure des mécanismes d'arrêt ou des mécanismes de renvoi du contrôle à l'utilisateur, manuel ou automatique, dans les systèmes éthiques dès la conception, puisqu'un tel renvoi suppose une compréhension des limites possibles du système et la pertinence d'une reprise en main par l'être humain. Enfin, si une telle approche du développement éthique est pertinente pour tout type de logiciel, elle est particulièrement valable pour les systèmes apprenants en continu, dans la mesure où la poursuite de l'apprentissage après le déploiement brouille la distinction classique entre développement et maintenance et mise à jour.

À ce stade, le lecteur pourrait craindre que l'inclusion de la vie du système sociotechnique ne fasse exploser les limites voulues de la portée de ce travail, et lui confère une ambition irréaliste. Il s'agit là d'un problème générique de la littérature sur l'éthique dès la conception. À mesure que se développe la littérature sur l'éthique de l'IA, en particulier dans l'approche large incluant l'ensemble du système sociotechnique, on peut craindre que la recherche mène à une accumulation d'attentes complètement irréalistes à l'égard des développeuses et développeurs. En plus d'être maitresses programmeuses, les développeuses d'IAs devront devenir économistes, juristes, sociologues et philosophes. Elles devraient faire face à une exigence à laquelle pratiquement aucun corps de métier ne doit se confronter, à savoir non seulement parvenir à effectuer un travail, mais anticiper toutes les interprétations et mésinterprétations possibles de leur travail et tâcher de les prévenir, on ne sait par quelle capacité miraculeuse. Elles devraient aussi être capables de trancher des dilemmes moraux avec une lucidité et un souci du bien commun dignes des plus grands sages. Une telle approche combinerait un caractère complètement irréaliste à un effet peu désirable de délégation complète de la réflexion éthique aux spécialistes de la technique. Il convient donc de distinguer une invitation saine à réfléchir aux conséquences du travail de développement dans les systèmes sociotechniques d'une attente irréaliste de résolution en amont de tous les problèmes imprévisibles que peut susciter l'introduction d'un nouveau système technique.

Il existe cependant une limite inhérente au processus de développement qui donne en retour une limite réaliste à la délimitation de notre objet, à savoir les difficultés de l'opérationnalisation elle-même. Face à l'accumulation d'attentes de plus en plus exigeantes à leur égard, un des

[15] Lignes directrices, note 9 p. 6.

principaux devoirs politiques et éthiques de la communauté scientifique est de communiquer avec honnêteté et clarté sur les limites de ses capacités, qu'elles soient contingentes ou définitives, relatives à des attentes en coût ou absolues. Face au paysage technologique extrêmement fluide de l'IA, et à l'absence de limites fondamentales démontrées pour bien des questions, une telle communication n'a bien sûr rien d'une évidence. La pression concurrentielle qui mène souvent les entreprises du numérique à survendre leur capacité, tout comme le sensationnalisme médiatique avide de prospectives catastrophistes ou radieuses, n'aident guère non plus à faire vivre un tel débat. C'est la raison pour laquelle l'accumulation d'expérience dans l'approche éthique dès la conception, et en particulier dans une approche respectueuse de la vie privée dès la conception, est essentielle pour nourrir le débat sur le sens, les effets secondaires et la portée de l'opérationnalisation de questions éthico-politiques. Les développeurs devront donc certes tâcher de prendre en compte l'intégralité du cycle de vie des systèmes sociotechniques, mais non pour phagocyter la réflexion éthique mais pour comprendre en profondeur l'approche d'éthique dès la conception.

Mais l'aberration du fantasme technocentriste n'est pas seulement due aux limitations des capacités planificatrices, ou aux limites de l'opérationnalisation. Elle est aussi due au fait que l'éthique dès la conception suppose souvent la définition de métriques de performance qui constituent en elle-même une décision éthique et politique. La définition d'une métrique de performance pour une valeur donnée constitue en soi un problème significatif, comme le montre les difficultés posées par la pluralité des définitions statistiques de l'équité[16]. Les discussions de l'affaire COMPAS, du nom du logiciel attribuant un score de risque de récidive à des détenus, en ont offert une illustration frappante : l'un des problèmes les plus profonds soulevés par cette affaire est que l'entreprise ayant développé ce logiciel a dû faire un choix de métrique de l'équité qui demeurait largement implicite pour les usagers et isolé du débat public. Or ce choix de métrique ne pouvait être considéré comme innocent dans la mesure où il existait dans ce cas un *trade-off* entre mesures de l'équité qui condamnait le logiciel à produire plus de faux positifs pour les détenus Afro-Américains[17]. À ces problèmes soulevés par la définition d'une valeur s'ajoute les problèmes posés par les relations entre les différentes valeurs poursuivies par le développement éthique, et dont les plus fréquentes sont l'équité et le respect de la vie privée. Il est courant de remarquer que les différentes valeurs poursuivies par les approches d'éthique dès la conception peuvent entrer en conflit. La charte 3.0 de l'*Association of Internet Researchers* (AoIR) souligne ainsi le conflit qu'il peut exister entre transparence, lutte contre les biais et contrôle par les sujets de données. Le respect de la vie privée pose aussi des problèmes pour la communication libre et ouverte des résultats de la recherche, essentiel à la reproductibilité. Comme le souligne également Auréalia Tàmo-Larrieux[18], il peut aussi exister un conflit entre sécurisation des données et interopérabilité des systèmes, le système le plus sûr n'étant pas conçu pour être interopérable avec des systèmes moins sûrs[19]. Tous

[16] Pour une introduction pédagogique à ces problèmes techniques, avec un commentaire de leurs implications en termes de philosophie politique, voir Reuben Binns, « Fairness in Machine Learning: Lessons from Political Philosophy », in *Conference on Fairness, Accountability and Transparency*, 2018, 149-59, http://proceedings.mlr.press/v81/binns18a/binns18a.pdf.

[17] L'article ayant lance la controverse est dû à l'association Propublica : Julia Angwin et Jeff Larson, « Machine Bias », text/html, ProPublica, 23 mai 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Pour une excellente introduction pédagogique au problème de choix de métrique, voir Karen Hao et Jonathan Stray, « Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game », *MIT Technology Review*, 17 octobre 2019, https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

[18] Aurelia Tamò-Larrieux, *Designing for Privacy and Its Legal Framework: Data Protection by Design and Default for the Internet of Things*, Issues in Privacy and Data Protection (Springer International Publishing, 2018), https://doi.org/10.1007/978-3-319-98624-1.

[19] Il convient aussi de souligner qu'elles peuvent aussi se renforcer mutuellement. Dans les *Lignes directrices en matière d'éthique pour une IA digne de confiance* du Groupe d'Experts Indépendants de Haut Niveau sur l'IA, GEHN IA mandaté par la Commission Européenne, il est proposé que l'utilisateur doit être informé des capacités et limitations du système dans le cadre de la quête de transparence. Parmi ces capacités et limitations figurent assurément les

ces exemples montrent que l'éthique dès la conception ne peut viser à une élimination du débat éthique par l'autorité de la science, car il n'existe pas de méthode scientifique pour arbitrer entre des valeurs incommensurables. Le résultat de cette approche ne doit pas être de clore le débat éthique sur les modèles, mais au contraire de l'initier. Elle ne mène pas à la suppression des dilemmes éthiques, mais force à effectuer des choix éthiques clairs, explicites et assumés. C'est précisément là que pourrait résider l'une des plus grandes vertus de cette approche, à savoir qu'elle force le concepteur du logiciel à sortir du bois éthique, et a déclaré très explicitement sa conception des problèmes suscités par son artefact et les solutions apportées, qui peuvent comprendre des arbitrages éthiques difficiles. Le fruit de cette approche ne sera pas tant une opérationnalisation des questions éthiques qu'une explicitation de la dimension éthique de la programmation.

Délimitation de l'objet et état actuel du droit : le problème de la finalité du traitement

Comme nous l'avons déjà évoqué, un autre problème structurant notre approche est la relation à l'état du droit : nous tâchons de nous concentrer sur des aspects du respect de la vie privée qui dépassent les enjeux déjà bien définis dans l'état de l'art juridique, en particulier pour l'encodage des données personnelles dans les modèles de ML. Nous voudrions expliciter notre position à l'égard d'un autre problème de l'état de l'art du droit européen, à savoir la définition de la finalité du traitement.

La définition de la finalité du traitement est essentielle au contrôle du sujet de données sur ces données personnelles. Dans le cas d'un consentement explicite, le sujet ne consent pas à n'importe quel traitement de ces données personnelles, mais à un traitement donné. La définition de la finalité du traitement permet sa restriction, qui est essentielle au respect de la vie privée. Nombre de problèmes de respect de la vie privée proviennent en effet du recoupement de bases de données auparavant séparées, ou de l'application d'un nouveau traitement sans rapport avec le traitement initial des données. En outre, la définition de la finalité du traitement est aussi importante pour la vérification des propriétés du modèle, dont nous parlerons plus bas (voir section II.A, *Les excès du pouvoir prédictif*) : la vérification ne peut s'opérer que sur le fond d'une spécification claire des finalités recherchées.

Dans le RGPD, la collecte des données personnelles est normalement limitée à des fins « déterminées, explicites et légitimes. » Il est cependant admis que la finalité de la recherche scientifique peut être plus difficile à délimiter : c'est même la seule exception au principe de détermination de la finalité du traitement. Si un changement de finalité de traitement doit être compatible avec la finalité initiale en général, dans la recherche cette compatibilité est présumée. Cela permettrait ainsi de récupérer des données auprès d'organismes les ayant collectées pour des fins différentes de la recherche[20]. Le droit européen contient aussi des exceptions sur les mesures de protection des données personnelles à l'endroit de la recherche. Selon le principe de conservation limitée, la conservation des données sous une forme permettant l'identification des personnes doit être limitée au strict minimum de temps nécessaire à la finalité du traitement, mais il y a là encore une exception pour la recherche. Le traitement de données sensibles et de données de santé (informations sur la santé physique et mentale des individus) est possible pour la recherche, même sans le consentement explicite des personnes concernées.

---

propriétés de sécurité et les dangers pour la vie privée, de sorte que transparence et respect de la vie privée peuvent se supporter mutuellement par le truchement de l'éducation de l'utilisateur.

[20] Ces remarques sont développées dans deux posts du blog juridique S.I.Lex, de Lionel Maurel *alias* calimaq : calimaq, « Données personnelles et recherche scientifique : quelle articulation dans le RGPD ? », - *S.I.Lex* - (blog), 18 juillet 2018, https://scinfolex.com/2018/07/18/donnees-personnelles-et-recherche-scientifique-quelle-articulation-dans-le-rgpd/. calimaq, « Affaire DisinfoLab : quelles retombées potentielles sur la recherche publique et la science ouverte ? », - *S.I.Lex* - (blog), 21 août 2018, https://scinfolex.com/2018/08/21/affaire-disinfolab-quelles-retombees-potentielles-sur-la-recherche-publique-et-la-science-ouverte/.

On est là à l'opposé du paradigme de pensée qui ferait de la recherche le parangon de l'explicitation préalable des hypothèses : la finalité de la recherche peut être élargie ou précisée en cours de recherche. Une telle exception est particulièrement pertinente pour la recherche en ML. Comme il a souvent été souligné, la recherche en ML fonctionne souvent dans un paradigme inductif, dont le mérite est précisément de laisser le modèle découvrir des régularités dans les données que le chercheur humain ne pouvait nullement anticiper. Une stricte détermination *ex ante* des finalités du traitement pourrait ainsi brider la dynamique propre à cette démarche scientifique.

Le droit européen accorde donc des libertés considérables aux chercheurs en matière de protection des données personnelles. Mais il existe toutefois un principe qui demeure plus problématique à cet égard. Selon le principe de minimisation, la collecte des données doit être limitée au strict nécessaire pour la finalité déclarée du traitement. Ce principe de minimisation s'applique tout aussi bien aux finalités de recherche. Il existe donc une tension interprétative entre le principe de minimisation, qui n'a de sens que si la finalité du traitement est clairement déclarée et ne peut être constamment modifiée au fil de l'eau, et les libertés accordées à la recherche en termes de définition de l'objet du traitement.

Une solution simple et naturelle consiste à préconiser une documentation rigoureuse de l'évolution de la finalité de traitement et de la collecte associée. À chaque étape de son cycle de développement, la déclaration de la finalité de son traitement doit être revue, et doit justifier scrupuleusement tout changement de finalité et toute extension de la collecte de données. Chaque modification de cette nature devrait s'accompagner d'une réflexion sur son potentiel impact sur le respect de la vie privée, par comparaison avec la première spécification de la finalité et de la collecte. Elle devrait aussi être justifiée par une authentique découverte scientifique modifiant les hypothèses justifiant la délimitation précédente des finalités et de l'extension de la collecte, et ouvrant de nouvelles possibilités dont l'intérêt scientifique doit être explicité. Cette solution est capturée dans notre principe 1 :

Dans le cadre de recherches scientifiques, déclarer les finalités de l'usage des données et l'extension de la collecte nécessaire à ces finalités, documenter et justifier tout écart à cette déclaration initiale par une découverte faite dans le cours de la recherche et l'intérêt scientifique d'une modification des hypothèses. Discuter explicitement les impacts possibles sur la vie privée de ces changements de finalité et de collecte.

Ce principe vise essentiellement à assurer un usage responsable de la liberté dérogatoire accordée aux chercheurs de dévier des finalités initiales de l'usage et d'étendre la collecte. Il impose une documentation rigoureuse des finalités et de leur évolution, ainsi qu'une justification proprement scientifique de cette évolution. Chaque nouvelle étape de cette évolution doit faire l'objet d'une discussion documentée des risques éventuels induits pour le respect de la vie privée, et des mesures prises pour mitiger ces nouvelles prises de risque.

Pour résumer, ce travail est centré sur l'approche éthique du développement des IAs, conçue pour inclure l'ensemble du cycle de vie du logiciel et non pas uniquement la conception *ex ante*. Nous nous concentrons sur les aspects de respect de la vie privée, en particulier le défi, propre aux modèles, posé par l'encodage de données personnelles dans les modèles eux-mêmes. Ce centrage est justifié à la fois par le caractère spécifique aux modèles dernière vague du ML de ce défi, et par l'insuffisance de sa conceptualisation dans l'état de l'art juridique (voir section III). Nous abordons aussi les problématiques liées aux définitions de la finalité du traitement. Là encore, cet intérêt est justifié à la fois par l'existence d'enjeux spécifiques au ML et par un certain flou du droit sur cette question, mais aussi parce qu'il est essentiel à dans la discussion de nombreux problèmes de respect de la vie privée, comme les limitations de la collecte, la limitation du pouvoir prédictif, la vérification des modèles et la limitation de la collecte (voir respectivement section II A, *Les excès du pouvoir prédictif*, II B *L'exclusion des données personnelles des bases d'entraînement* et *L'approche de sécurité* post hoc).

II A. Failles de sécurité des modèles et respect de la vie privée : un état de l'art raisonné aux fins de discussion éthique

Dans cette deuxième partie, nous offrons un bref état de l'art des problèmes de sécurité liés aux données personnelles dans le ML, avant d'aborder dans la section suivante les différentes réponses techniques offertes à ces problèmes. Si un tel objectif implique évidemment de rentrer dans le détail des techniques, notre objectif n'est pas tant de faire un état de l'art complet que de donner les informations suffisantes à la position des problèmes éthiques soulevés par la configuration technique présente. C'est la raison pour laquelle nous restons au stade d'une présentation intuitive, et référons le lecteur à la littérature nécessaire à une compréhension intime des problèmes techniques. Certaines des attaques que nous allons présenter peuvent être effectuées sans accès au code du modèle, par une simple analyse de son comportement entrées-sortie : on parle alors d'attaque en « boite noire ». Pour les attaques supposant un accès aux rouages du modèle, on parle d'attaque en « boite blanche ».

L'inversion des modèles

Le premier problème de sécurité posé par les modèles d'apprentissage automatique est celui d'inversion des modèles de ML. Pour le résumer en une formule simple, ces modèles enregistrent pendant leur phase d'apprentissage certaines informations sur leur base de données d'entraînement, informations qui peuvent être ensuite récupérées, même si la base de données initiale a été détruite, en attaquant le modèle lui-même. Cette récupération d'informations, notamment dans le cas des bases de données textuelles, peut aller jusqu'à la récupération de points de données, pouvant contenir des données personnelles. Certains modèles entraînés sur des corpus peuvent même contenir une quasi-copie de leur base de données d'entraînement, qu'ils ont pour ainsi dire appris par cœur : une attaque sur ces modèles permettrait donc une récupération de l'essentiel des informations contenues dans le jeu d'apprentissage. Toutes ces failles constituent un enjeu de sécurité majeur pour tout modèle entraîné sur des bases contenant des informations personnelles.

Du point de vue de la culture de la sécurité informatique, le grand problème posé par ces attaques est qu'il n'existe pas à l'heure actuelle de résultats durs permettant d'ériger des barrières de sécurité mathématiquement prouvées, comme cela peut être le cas en cryptographie. Les différentes approches existantes permettent seulement de bloquer certaines attaques, ou de diminuer fortement leurs performances[21]. Dans un tel état de l'art, il n'existe donc pas de solution technique directe au problème des attaques par inversion, et l'on se doit donc de chercher soit des solutions techniques de contournement soit des solutions extratechniques.

*Les attaques par inférence d'appartenance (***membership inference***)*

Ce deuxième type d'attaque ne vise pas tant à inverser le processus d'entraînement pour retrouver les données à partir du modèle, qu'à inférer si un sujet de données *S* fait partie de l'ensemble de données d'entraînement, avec éventuellement l'aide d'un autre ensemble de données contenant le sujet *S*. Comme l'inversion de modèles, l'inférence d'appartenance peut être effectuée en boite noire ou en boite blanche[22].

La simple appartenance d'un sujet de données à une base peut évidemment être une information très sensible : il suffit de penser à l'appartenance à une liste de patients psychiatriques ou à une liste de délinquants récidivistes. Si elles révèlent potentiellement moins d'informations

[21] Yasmeen Alufaisan, Murat Kantarcioglu, et Yan Zhou, « Robust Transparency Against Model Inversion Attacks », *IEEE Transactions on Dependable and Secure Computing*, 2020, 1-1, https://doi.org/10.1109/TDSC.2020.3019508.

[22] Shilin Qiu et al., « Review of artificial intelligence adversarial attack and defense technologies », *Applied Sciences* 9, nº 5 (2019): 909.

que les attaques par inversion de modèles, les attaques par inférence d'appartenance peuvent révéler une information tout aussi sensible.

Les excès du pouvoir prédictif

Un autre problème est dû à l'excès de pouvoir prédictif de certains modèles, qui permet de récupérer des données personnelles alors même que cette fonction ne fait pas partie de la spécification du logiciel. Un bon exemple d'un tel problème a été analysé dans l'étude d'un logiciel de complétion[23]. Ce logiciel a pour spécification de formuler des suggestions de mots ou d'expressions à l'utilisateur : son objectif est donc d'apprendre des traits génériques de la pratique de la langue écrite, et non d'apprendre des données personnelles de l'utilisateur. Cependant, l'apprentissage sur la pratique écrite de l'utilisateur permet au logiciel d'apprendre des données à grain plus fin, qui peuvent capturer des données personnelles. Par exemple, en tapant le début de phrase « mon numéro de carte de crédit est… », l'utilisateur peut se voir suggérer un numéro de carte exact. Même si l'utilisateur a détruit les fichiers contenant ces données, le modèle permet donc leur récupération en boite noire par une attaque extrêmement simple, qui consiste à taper un début de phrase appelant une complétion contenant l'information.

Un tel problème arrive tôt dans l'apprentissage et n'est donc pas le fruit d'une suroptimisation. Il persiste à travers plusieurs modèles et apparaît même pour des modèles de petite taille par rapport au corpus. Ce dernier point est important car il exclut une pure assimilation « par cœur » des données : il montre que c'est bien le pouvoir prédictif du modèle qui est problématique.

On voit donc que la faille de sécurité provient d'un excès de pouvoir prédictif du modèle : le modèle a appris des données trop fines, qui ne font pas partie de sa spécification initiale et qui créent une faille de sécurité. Une solution possible serait donc de modifier l'entraînement du modèle afin de réduire son pouvoir prédictif à des données à plus gros grain, qui ne pourraient contenir d'informations personnelles par dessein. Nous verrons que c'est ce qu'essaye en partie de réaliser de manière probabiliste les approches de confidentialité différentielle (*differential privacy*), en injectant du bruit dans le processus d'apprentissage (voir section éponyme). Tout ceci démontre la grande importance d'une définition fine des objectifs d'apprentissage pour éviter de telles surprises. Le développeur doit donc travailler non seulement pour maximiser le pouvoir prédictif de son modèle, mais aussi pour le restreindre à ce qui est strictement nécessaire à la complétion de la tâche considérée. Si cela peut être vu comme une perte de puissance scientifique, cela peut être aussi vu comme un appel à une connaissance plus fine, qui évite la restitution de points de données pour capturer des traits plus abstraits et plus génériques.

Les développeuses et développeurs doivent donc tester les performances prédictives de leur modèle pour détecter un éventuel pouvoir prédictif à la fois dangereux pour la vie privée et excessif par rapport aux objectifs d'apprentissage et aux finalités de traitement. Elles doivent ensuite prendre toutes les mesures possibles afin de restreindre le pouvoir prédictif à ce qui est à la fois nécessaire et vertueux, que ce soit par exemple en modifiant l'apprentissage ou en filtrant des résultats. Cette lutte contre les excès du pouvoir prédictif est capturée dans notre principe 2 :

Afin de prévenir l'apparition d'un pouvoir de prédiction trop fin présentant une menace pour la vie privée, tester les performances finales du logiciel en s'interrogeant sur la nécessité des résultats par rapport aux objectifs d'apprentissage et aux finalités du traitement. Prendre toutes les mesures possibles pour limiter le pouvoir prédictif du modèle à ce qui est strictement nécessaire pour ces objectifs et finalités.

Les développeuses et scientifiques sont plutôt habituées à travailler d'arrache-pied à améliorer les performances prédictives de leur modèle. Ce principe les invite donc à un exercice inhabituel, à savoir la méfiance par rapport à des performances prédictives excessives par rapport

[23] Nicholas Carlini et al., « The secret sharer: Measuring unintended neural network memorization & extracting secrets », 2018.

à la finalité du traitement. Comme pour tout exercice inhabituel, une accumulation d'expérience et une réflexion méthodologique concomitante seront nécessaires pour garantir sa bonne exécution par toutes et tous. Les réflexions méthodologiques devront en particulier porter sur les outils permettant de limiter la granularité de l'apprentissage ou de désapprendre un savoir devenu excessivement précis. L'intérêt scientifique de ces outils est qu'ils pourraient permettre l'émergence d'un apprentissage plus fin et d'une meilleure connaissance réflexive des modèles, soit une opportunité de faire de nécessité éthique vertu scientifique.

La lutte contre la suroptimisation

Une autre source de pouvoir prédictif excessif, et donc dangereux pour la vie privée, est le phénomène de suroptimisation : au lieu d'apprendre un signal généralisable, le modèle peut se mettre à apprendre « par cœur » les traits particuliers de sa base d'entraînement. Si cette base comprend des données personnelles, il risque de restituer dans ces prédictions des données personnelles très fines, et donc de devenir invasif. La communauté de ML étudie depuis longtemps ce phénomène et des mesures de vérification existent, mais encore une fois sans résultat absolument certain, surtout avec les nouveaux modèles d'apprentissage profond (*Deep Learning,* DL), qui possèdent plus de paramètres que de données d'apprentissage. L'un des grands intérêts de cette lutte pour la vie privée dès la conception est de stimuler des recherches méthodologiques sur la lutte contre la suroptimisation bénéfiques pour la communauté entière du ML. Comme pour l'excès de pouvoir prédictif, les mesures de protection de la vie privée dès la conception ne doivent pas être vues comme une restriction de la connaissance scientifique, mais comme un appel à une connaissance scientifique plus fine.

Néanmoins, de telles mesures ne procurent pas une sécurisation complète du modèle, puisque des attaques comme les attaques d'inversion peuvent réussir même en l'absence de suroptimisation. Enfin, et surtout, la lutte contre la suroptimisation est un problème structurel du ML. Chaque développeuse et développeur connait fort bien le problème de la suroptimisation : les inviter à lutter contre ce phénomène peut donc sembler aussi creux que de les appeler à développer de bons modèles. Néanmoins, il est bon de signaler que la suroptimisation est aussi un éventuel problème pour le respect de la vie privée dès la conception. Il sera bon d'examiner dans les cas d'usage particuliers si la lutte contre la suroptimisation est une voie particulièrement praticable, ou au contraire un choix difficile à assumer. Dans tous les cas de figure, il est bon que le respect de la vie privée soit inclus dans la réflexion sur la suroptimisation, en particulier lorsqu'un arbitrage entre suroptimisation et pertes de performances doit être effectué lors de la terminaison de l'apprentissage. Cette inclusion est résumée dans le principe 3 :

Dans les arbitrages sur la terminaison du processus d'apprentissage, et le compromis à trouver entre risque de suroptimisation et perte de performances, prendre en compte les risques posés par la suroptimisation pour la vie privée.

Ce principe est bien distinct du principe 2, dans la mesure où ce dernier visait à affronter des problèmes d'excès du pouvoir prédictif qui ne sont pas dus à la suroptimisation. À la discussion scientifique usuelle sur le Charybde de la sousoptimisation et le Sylla de la suroptimisation, il injecte un principe éthique extrinsèque qui, s'il n'influence en rien la discussion méthodologique de fond sur l'apprentissage, peut influencer les arbitrages concrets pris par les développeurs en faveur de la vie privée.

Le pouvoir d'inférer des données personnelles à partir de données publiques

La troisième évolution technologique, qui présente un problème majeur pour le respect des données personnelles, est la capacité d'inférence statistique permettant de déduire des données personnelles à partir de données qui ne seraient pas considérées comme problématiques, soit parce qu'elles sont anonymisées, soit parce qu'elles sont publiques. Puisque les modèles d'apprentissage ont démontré des capacités, parfois spectaculaires, pour détecter des corrélations entre différents

ensembles de données, ils peuvent trouver des corrélations importantes entre de telles données et des données personnelles parfois problématiques qui peuvent servir de base à une inférence statistique déduisant les secondes à partir des premières. Ce pouvoir d'inférence statistique constitue une remise en cause fondamental du paradigme de protection des données personnelles par restriction de l'accès à ces données : rien ne sert d'enfermer des données personnelles dans le meilleur des coffres forts numériques si l'on peut les deviner en s'appuyant sur d'autres données aisément accessibles. Lorsqu'une une inférence statistique d'une donnée personnelle à partir d'une autre donnée est possible, même la restriction la plus ultime de l'accès, à savoir la destruction, ne peut empêcher la récupération de données problématiques à partir du modèle. Les plus fortes mesures de sécurisation des bases de données peuvent donc devenir inefficientes si certaines informations contenues dans ces bases peuvent faire l'objet d'une inférence statistique. L'apprentissage automatique devient donc une faille de sécurité majeure pour la protection des données, qui demande une réflexion technique et normative dédiée.

On pourrait objecter que la faculté de déduire des données personnelles, y compris à partir d'inférence statistique, n'a rien de nouveau. Cependant, l'expansion fulgurante de la collecte et de l'apprentissage, en plus d'une considérable augmentation de ce pouvoir d'inférence, créent une incertitude et une fluidité du champ qui est nouvelle. Il devient difficile d'anticiper quelles données sont en sécurité, et quelles sont soumis à des menaces, dans un monde où l'analyse de données crée sans cesse de nouvelles capacités d'inférence.

Il existe bien sûr toujours des données personnelles qui ne seront pas inférables à partir d'autres données, et dont la protection représente encore un enjeu majeur. Mais reste à savoir si une telle approche peut toujours passer à l'échelle, ou si elle est condamnée à devenir une approche minoritaire dans un monde où la plupart des données personnelles seront inférables. Les évolutions récentes n'incitent certainement pas à l'optimisme sur les capacités techniques à se protéger contre des inférences invasives, mais une étude devrait encore être faite pour savoir combien de nos données sont encore protégées contre de telles inférences, et combien sont d'ores et déjà aisément inférables.

Cette question des limitations des capacités techniques à lutter contre des inférences invasives est également décisive pour toute approche de développement éthique respectueuse de la vie privée. Une telle approche n'est évidemment pas sans pertinence pour la lutte contre les inférences invasives. Comme nous le verrons ci-dessous, lutter contre la suroptimisation des modèles peut permettre de limiter leur pouvoir d'inférence à des données génériques sans incidence sur la vie privée. Mais il reste à voir si de telles approches de limitation des capacités d'inférence peuvent passer à l'échelle, ou si l'essentiel de la lutte contre les inférences invasives doit être centrée soit sur des restrictions sur l'usage, soit sur des restrictions sur les données qu'il sera légal de collecter, conserver et exploiter. Cette dernière approche peut aussi être considérée comme une approche de respect de la vie privée dès la conception, dans la mesure où la collecte des données fait bien partie du travail du développement, mais elle se démarque nettement d'une restriction sur la conception des modèles eux-mêmes. La possibilité des inférences invasives à partir de données non-personnelles constitue donc un problème décisif pour la forme à adopter et la pertinence de l'approche de respect de la vie privée dès la conception.

II B Solutions possibles aux problèmes de respect de la vie privée

Pour un entraînement des modèles respectueux de la vie privée

La première solution envisageable pour le problème d'inversion des modèles serait de restreindre l'accès à ces modèles. Cette approche en boite noire serait cependant antithétique de la culture de publicité du logiciel, en créant une justification facile à la privatisation et au secret sous couvert de sécurité. Quelles sont les autres options techniques ?

1. *L'exclusion des données personnelles des bases d'entraînement : de la définition des données problématiques à l'exclusion des données personnelles de d'apprentissage*

Une réponse simple dans son principe, mais puissamment contraignante dans la pratique, serait de s'interdire tout emploi de données personnelles dans l'entraînement des modèles. Puisqu'en l'état actuel de l'art en sécurité de l'apprentissage automatique, on est incapable d'empêcher complètement la récupération des données personnelles à partir des modèles, la création de contraintes sur l'entraînement permettrait de concilier respect de la vie privée et publicité du logiciel[24].

Cette proposition constituerait une rupture radicale avec les pratiques en cours[25], et rencontrerait probablement une résistance considérable non seulement des développeuses, entreprises et autres institutions productrices de modèles, mais même de certains utilisatrices qui souhaitent être vues, référencées, suggérées, etc. Mais elle constitue une question à se poser pour les développeuses et développeurs : puis-je entraîner mon modèle sans données problématiques ? Il faut commencer à se poser cette question pour comprendre le rôle joué par le traitement des données personnelles dans les pratiques actuelles, et estimer l'impact et le sens qu'aurait une telle restriction sur le développement.

Une critique possible de cette approche serait qu'elle est d'une force excessive. D'après le considérant 26 du RGPD et l'article 4-1, compte comme donnée personnelle « toute information concernant (*relating to*) une personne physique identifiée ou identifiable. » L'identification ne se réduit pas à la possession du nom de la personne, ou identification directe. Elle désigne de manière plus générale la capacité à singulariser une personne au sein d'une population : lorsque qu'une combinaison de propriétés qui ne sont pas identifiantes en elles-mêmes permettent de singulariser une personne physique au sein d'une population, on parle d'identification indirecte. La définition des données personnelles ne fait pas référence à la notion de vie privée : une information personnelle peut parfaitement être publique et ne pas porter atteinte à la vie privée. La notion d'information est comprise dans un sens très large, sans prise en compte de sa nature : une information personnelle peut être vraie ou fausse, précise ou imprécise, subjective ou objective, et comprendre des opinions et des estimations. Le contenu de l'information n'est nullement restreinte aux informations sur la vie privée ou familiale, dans la mesure où la protection des données vise à protéger les droits des individus, qui comprennent mais ne sont pas réduits à la vie privée[26]. Par contraste, les données anonymes sont soit les données ne faisant nullement référence à une

[24] Sans aller jusqu'à envisager l'interdiction, le GEHN IA envisage une telle approche (*Lignes directrices,* p.36), en posant cette question aux développeurs : « Avez-vous réfléchi à des manières de mettre au point le système d'IA ou d'entraîner le modèle sans utiliser (ou en utilisant de manière limitée) des données potentiellement sensibles ou à caractère personnel ? »

[25] Il a ainsi déjà été proposé d'exclure des modèles de profilage les « informations pertinentes pour une personne » (« *personnally relevant informations* »). Il s'agit là d'une notion plus ample que les notions de donnée personnelle ou de donnée sensible, et faite pour prendre en compte les problèmes posés par les capacités d'inférence. Les informations pertinentes pour une personne sont toutes les informations permettant de déduire des informations sensibles sur la personne. Karina Vold et Jessica Whittlestone, « Privacy, Autonomy, and Personalised Targeting: rethinking how personal data is used », 2019.

[26] De manière générale, la notion de donnée personnelle est complètement indépendante de la nature et du contenu de l'information considérée, comme du médium par lequel elle est transmise. Elle n'a pas besoin d'être vraie ou objective, pas plus qu'elle n'a besoin d'être secrète ou privée. Elle doit juste faire référence à une personne physique identifiable : *« Any kind of information, regardless of its nature, content, format or the medium in which it is contained, can qualify as personal. It does not need to be truthful or objective, nor secret or private, nor kept in a particular format or medium. Any sort of data can be personal, if it relates to an identifiable natural person »* Lorenzo Dalla Corte, « Scoping Personal Data: Towards a Nuanced Interpretation of the Material Scope of EU Data Protecton Law », *European Journal of Law and Technology* 10, nº 1 (16 mai 2019), http://ejlt.org/index.php/ejlt/article/view/672.

personne physique, soit celles traitées de manière à ce que les personnes physiques ne soient plus identifiables.

Cette notion est aussi bien plus ample que la notion de « donnée sensible », c'est-à-dire les données touchant à des informations cruciales sur la personne, pouvant notamment être utilisées à des fins discriminatoires. Les données sensibles sont dûment énumérées dans l'article 9 du RGPD, et contiennent le genre, l'appartenance ethnique, la confession, l'orientation sexuelle, l'état de santé, les données biométriques ou génétiques permettant l'identification d'une personne physique, l'appartenance à un syndicat ou un parti politique, et les opinions politiques, religieuses ou philosophiques en général. Ces données bénéficient d'un régime de protection renforcée, avec notamment une interdiction générique de traitement. Il convient de remarquer en passant que le législateur a aussi pris en compte la possibilité offerte par le ML d'inférer des données sensibles à partir d'autres données collectées par des tierces parties, comme l'historique d'achat sur un site commercial ou les diverses activités sur un réseau social. Plusieurs travaux ont montré la possibilité d'inférer avec une forte probabilité non seulement des données personnelles, mais même des données sensibles à partir de ce type de données, comme l'état de santé psychique, la confession, les opinions politiques, l'orientation sexuelle ou l'origine ethnique[27],[28],[29],[30]. Le RGPD impose de classer ces données comme des données sensibles à partir du moment où elles permettent de telles inférences, ce qui signifie que certains modèles de ML, parce qu'ils encodent de telles données, pourront être considérées comme relevant du régime des données sensibles[31] (voir section III.1). Comme le remarque à juste titre A. Drozdz, l'explosion des corrélations entre données personnelles « ordinaires » et données personnelles sensibles, notamment par la pratique du profilage, vient brouiller la séparation entre les deux catégories. Non seulement « tout est donnée personnelle », mais « tout est donnée est sensible » (voir section III. 3 pour une discussion de ce problème). Par conséquent, il existe un risque non-négligeable que nombre de modèles de ML finissent par relever de la législation sur les données sensibles, ou à tout le moins posent problème à cette législation.

Éliminer toute donné personnelle impliquerait donc d'éliminer toute référence aux personnes physiques (ou à des groupes restreints de personnes physiques). Si l'on a affaire à des ensembles de données mélangeant données personnelles et données anonymes, ceci pose déjà des questions importantes sur les performances des outils de reconnaissance d'entités dans les images ou dans un corpus. De manière encore plus problématique, cela pourrait impliquer d'être capable d'identifier une référence à une personne physique qui n'est ni nommée, ni désignée par un titre, ni représentée explicitement dans une image.

L'élimination de toute référence à une personne physique pourrait éliminer beaucoup d'informations publiques sur les personnes physiques qui n'ont rien de problématique : l'élection à une position de responsabilité, l'acquisition d'une compagnie, la prise de position dans un débat public, la sortie d'un album… contiennent toutes des références à des personnes physiques, et sont donc des données personnelles au titre du droit. Toutes ces informations sont non seulement publiques par nature, mais le sujet de données peut même souhaiter la diffusion de cette information et y contribuer activement.

Si on souhaite éviter de tailler trop large, on pourrait donc restreindre l'approche pour exclure de sa portée les données personnelles dont la diffusion ne pose pas de problèmes de respect de la

---

[27] *View of Gaydar: Facebook friendships expose sexual orientation \textbar First Monday*, consulté le 11 novembre 2020, https://firstmonday.org/article/view/2611/2302.

[28] Michael Reilly, *Is Facebook targeting advertising at depressed teens* (MIT Technology review, 2017).

[29] Jeremy B. Merrill, « Liberal, moderate or conservative? see how facebook labels you », *The New York Times* 23 (2016).

[30] Annalee Newitz, « Facebook's ad platform now guesses at your race based on your behavior », *Ars Technica* 18 (2016).

[31] Voir Aleksandra Drozdz, *Protection of Natural Persons with Regard to Automated Individual Decision-Making in the GDPR,* Kluwer Law International BV, 2020, ainsi que les références mentionnées dans les notes 41 à 44.

vie privée[32] ou d'autres droits de la personne. Trouver un tel critère de démarcation entre données personnelles à protéger et données personnelles non-problématiques constitue donc un problème redoutable d'opérationnalisation de l'approche. Cette approche bute ainsi sur un problème à la fois classique et fondamental des approches vertueuses dès la conception, à savoir la difficulté à opérationnaliser les vertus. Le droit est coutumier de l'emploi de notions vagues, qui ne reçoivent qu'une définition provisoire, souvent accompagnée par une liste d'exemples privilégiés : il revient ensuite aux juges de préciser cette définition par une interprétation au cas par cas. Une telle approche n'est pas envisageable en l'état en informatique, où l'on a besoin d'une définition des notions employées en amont. Le geste d'une telle définition pose deux problèmes essentiels. Le premier est épistémique : la définition de la notion de « donnée personnelle dont la diffusion porterait atteinte aux droits de la personne, en particulier à sa vie privée » est une définition redoutablement difficile à obtenir, et sa possibilité même peut faire débat. Le second est d'ordre politique et juridique : en donnant une définition que ni les juristes ni le débat politique n'ont réussi à donner, les informaticiennes et informaticiens se donnent un grand pouvoir politique, dont la légitimation institutionnelle fera certainement débat : même si leur travail arrivait à une définition digne d'intérêt, il leur incombera encore de convaincre la communauté des citoyens et des juristes de la pertinence d'une telle définition pour qu'elle puisse véritablement devenir une norme.

Une variante de l'approche par restriction des données d'entraînement consisterait non pas à chercher *la* définition de ce type de données, mais *une* définition de cette notion, de préférence celle adoptée par le pays où opèrent les développeuses et développeurs. Une telle approche pourrait parfaitement être fondée sur une philosophie rejetant toute tentative de définition scientifique d'une telle notion comme un pur mirage positiviste, qui ne saurait atteindre aucun résultat substantiel. La seule approche raisonnable d'une telle tâche serait d'opérationnaliser *une* définition juridique ou philosophique existante. Une telle approche n'échappe bien sûr pas aux difficultés redoutables posées par l'opérationnalisation de tels concepts, puisque reconnaître l'absence d'unicité de la solution du problème ne préjuge pas de la faisabilité de l'implémentation d'une solution donnée. Elle a cependant l'avantage de poser moins de problèmes de légitimité politique, dans la mesure où les informaticiennes et informaticiens s'y cantonnent à une fonction instrumentale d'implémentation d'une notion préexistante, qui pourrait déjà faire l'objet d'un consensus dans la communauté politique considérée. La tâche ne consiste moins alors à définir une notion vague qu'à formaliser une définition possible d'une telle notion. Même sous cette forme faible, la possibilité d'opérationnaliser la notion pertinente de donnée problématique reste en l'état actuel très brumeuse et lointaine.

Le ML offre cependant une autre modalité d'opérationnalisation de notions, qui ne passe pas par leur définition explicite. D'un point de vue pratique, le problème essentiel est d'être capable de détecter des informations personnelles voire sensibles dans de vastes bases de données dont la revue manuelle serait très coûteuse ou même franchement impossible[33]. Lorsqu'on dispose de grandes quantités de données, le ML offre justement la possibilité de détecter la présence d'une entité sans jamais fournir de définition de cette entité. Une telle approche a déjà été appliquée aux données personnelles, qu'il s'agisse de données purement textuelles[34], ou de données textuelles

[32] Une telle approche peut naturellement être généralisée à d'autres problèmes moraux pris en charge par l'approche éthique dès la conception, comme la lutte contre les usages discriminatoires d'un modèle. Mais il s'agit là de sujets pour d'autres travaux dédiés.

[33] À titre de remarque, nous ne connaissons pas de travaux ayant essayé de procéder par passage à l'ensemble complémentaire, soit en identifiant les données personnelles qui ne posent pas a priori de problèmes légaux de diffusion, comme la participation d'une personne à un événement public, et considérer par défaut que les autres données personnelles doivent être purgées.

[34] « (PDF) AntShield: On-Device Detection of Personal Information Exposure », ResearchGate, consulté le 24 septembre 2020, https://www.researchgate.net/publication/323570796_AntShield_On-Device_Detection_of_Personal_Information_Exposure. « Privacy Disclosures Detection in Natural-Language Text Through Linguistically-Motivated Artificial Neural Networks | Request PDF », ResearchGate, consulté le 24

insérées dans des images ou vidéos[35], ou des images[36]. Il existe ainsi des modèles permettant la détection d'une référence à une personne physique (*Personally Identifiable Information,* PII) dans des e-mails, la fuite de données personnelles dans le trafic de données (*PII privacy leakage in data traffic*), des réseaux de neurones classant des informations textuelles en sensible, personnelle et non-personnelle.

Dans la pratique, l'absence d'une définition explicite ne signifie cependant pas l'absence de toute hypothèse sur la notion d'information privée ou personnelle. La constitution de la base de données, et les instructions données aux personnes annotant les corpus d'entraînement, imposent de formuler de telles hypothèses. Ravazi et Ghazinour[37] élargissent la notion de *Personal Health Information*[38] (PHI) pour inclure les adresses, noms, localisations, dates définissant l'âge, URLs, et les informations proprement médicales comme les médicaments, les noms de maladie, les symptômes, les assureurs tout comme les examens médicaux et les résultats. La PII[39] a une définition restreinte comme les informations monétaires, les adresses physiques et mails, et les numéros de téléphone. Le PrivacyBot[40] a cherché à étendre les approches précédentes en labellisant les informations de la base d'entraînement selon les catégories employées par le RGPD. D'après ces catégories, en plus de la liste d'informations sensibles citée plus haut, les informations personnelles incluent le nom, les e-mails et autres identifiants en ligne, les informations spatiales comme la localisation, l'adresse personnelle ou de travail, les numéros de sécurité sociale ou les codes bancaires, le numéro de téléphone, la profession, le lieu de travail, l'éducation, mais aussi des catégories plus vagues comme les données physiologiques ou sociales. On voit que toutes ces approches, parfois spécialisées pour une tâche précise, sont très variées et sont fondées, comme il est fréquent dans la pratique des définitions juridiques, sur des listes non-exhaustives. En revanche, tandis que l'approche juridique permet l'emploi ultérieur d'un juge humain pour trancher un cas difficile, les approches par ML mentionnées excluent un tel jugement *post hoc*, et ne capturent donc qu'une partie de la pratique juridique de définition et de qualification d'un cas. Elles sont réduites à une approche *ex ante* et centrée sur les listes définies à l'avance, où l'intuition humaine est limitée aux jugements des annotateurs de la base d'entraînement. L'une des grandes questions ouvertes du domaine est de voir à quel point une telle approche fondée sur une liste d'exemples privilégiés

---

septembre 2020, https://doi.org/10.1007/978-3-030-21373-2_14. Qiwei Jia et al., « Who Leaks My Privacy: Towards Automatic and Association Detection with GDPR Compliance », in *Wireless Algorithms, Systems, and Applications*, éd. par Edoardo S. Biagioni, Yao Zheng, et Siyao Cheng, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2019), 137-48, https://doi.org/10.1007/978-3-030-23597-0_11.Christoph Bier et Jonas Prior, « Detection and Labeling of Personal Identifiable Information in E-Mails », in *ICT Systems Security and Privacy Protection*, éd. par Nora Cuppens-Boulahia et al., vol. 428, IFIP Advances in Information and Communication Technology (Berlin, Heidelberg: Springer Berlin Heidelberg, 2014), 351−58, https://doi.org/10.1007/978-3-642-55415-5_29.Jiaqi Wu, « An Automated Privacy Information Detection Approach For Online Social Media » (PhD Thesis, Auckland University of Technology, 2019).

[35] « Detection of Sensitive Textual Information in User Photo Albums on Mobile Devices - IEEE Conference Publication », consulté le 24 septembre 2020, https://ieeexplore.ieee.org/abstract/document/8958325.

[36] Rui Jiao, Lan Zhang, et Anran Li, « IEye: Personalized Image Privacy Detection », in *2020 6th International Conference on Big Data Computing and Communications (BIGCOM)*, 2020, 91−95, https://doi.org/10.1109/BigCom51056.2020.00020.

[37]A.H. Ravazi and K. Ghazinour, « Personal health information detection in unstructured webdocuments », *IEEE 26th Symposium on Computer-Based Medical Systems* (CBMS), 2013, pages 155-160.

[38]La notion de PHI, et son exploitation par désidentification, étaient déjà utilisées avant la dernière vague du ML. Voir par exemple Khaled El Emam, *Guide to the de-identification of personal health information*. CRC Press, 2013.

[39]Al-Fedaghi, S. A., & Thalheim, B. (2008, November). Databases of personal identifiable information. In *2008 IEEE International Conference on Signal Image Technology and Internet Based Systems* (pp. 617-624). IEEE.

[40]Welderufael B. Teslay, Jetzabel Serna, Kai Rannenberg, « PrivacyBot : Detecting Privacy Sensitive Information in Unstructured Texts », International Conference on Network Analysis, Management and Security (SNAM), 2019

limite les capacités des modèles, dans la mesure où la méthodologie du ML est fondée sur une capacité de généralisation à partir des exemples présents dans la base d'entraînement.

Le problème posé par la définition de la notion employée ne peut qu'être renforcé par l'extension de cette approche des données privées ou sensibles à l'ensemble des données personnelles. Comme nous venons de le voir, la notion de donnée personnelle est beaucoup plus vaste que celle de données touchant à la vie privée. Capturer la notion suppose de capturer la référence à une personne physique sous toutes ses formes. Un simple coup d'œil à l'immense littérature en linguistique et philosophie du langage sur la notion de référence suffit à faire anticiper la difficulté d'un tel objectif. Qu'on songe aussi à la pratique courante dans les œuvres littéraires dites « à clef » consistant à faire référence à une personne sans la nommer ou en la pseudonymisant, mais en donnant des détails suffisamment précis pour qu'ils permettent une ré-identification à tout lecteur possédant suffisamment d'éléments de contexte. Dans la vie quotidienne, des allusions contextuelles comme « la personne dont nous avons beaucoup parlé a de gros problèmes de santé » sont aussi courantes. De tels subterfuges sont aisés à produire et à comprendre pour des êtres humains, mais sont voués à constituer un grand défi pour les programmes de reconnaissance d'informations personnelles, comme ils ont d'ailleurs souvent tenu en échec les autorités judiciaires. Non seulement les différentes notions employées ont des extensions différentes, mais elles pourraient aussi poser des problèmes techniques différents.

Les résultats des premiers travaux en détection d'information personnelle ou privée ne sont pas décourageants. Les études réalisées sont cependant limitées à un certain type de données, comme les e-mails professionnels ou les échanges sur les réseaux, et donc à un certain public d'utilisateurs et un certain nombre de pratiques affectant la vie privée. Elles sont aussi limitées par la langue, la taille et le type des corpus considérés, ou les hypothèses faites sur l'information privée utilisée. La nature statistique de l'approche est une autre limite structurelle. Par exemple, le réseau de neurones réalisant la classification tripartite non-personnel/personnel-sensible a une performance de 89%. Si une telle performance peut déjà grandement contribuer à réduire la charge de travail, surtout lorsqu'elle est accompagnée d'un écart de confiance guidant la vérification manuelle, elle n'en reste pas moins susceptible de laisser fuiter des informations cruciales.

On a donc affaire à un domaine encore jeune et peu structuré, même si les évolutions de la législation et la conscience grandissante des problèmes de vie privée numérique pourraient modifier cet état de fait à court terme. Quel que soit le futur de ce domaine de recherche, il est intéressant de remarquer dès à présent que la possibilité de purger automatiquement une base de données de l'écrasante majorité des informations touchant à la vie privée est ouverte, et que cela pourrait représenter une vraie possibilité d'implémentation pour l'approche basée sur l'exclusion des données personnelles.

Une autre sous-approche de la restriction des données d'entraînement, l'approche par restriction aux données explicitement publiées, pourrait offrir une solution de contournement des difficultés de l'exclusion des données que nous venons de mentionner. L'approche centrée sur le geste de publication part d'un postulat implicite sur la notion de vie privée, à savoir que le sujet de données est maître de l'accès à ces données privées : libre à lui de rendre public les détails les plus intimes de son existence. Ce qui compte n'est donc pas tant la nature sensible des données que la décision du sujet de les rendre public ou non. On opère donc un glissement de la notion de donnée personnelle dont la diffusion pourrait être problématique à celle de donnée publiée : une donnée personnelle publiée est toujours une donnée personnelle, mais y accéder n'est plus problématique si le sujet de données a fait usage de son droit de publication, c'est-à-dire s'il a donné accès à tous à certaines de ses données personnelles.

On pourrait donc prendre comme référence le geste de publication, et éviter tout document faisant référence à une ou des personnes physiques qui n'est pas accompagné d'une décision de publication explicite ou d'une licence ouverte. Il s'agirait là encore d'une rupture radicale avec la pratique actuelle, puisqu'elle reviendrait essentiellement à abandonner la mine de données que

constituent, notamment pour le traitement automatique de la langue et la reconnaissance d'images, les réseaux sociaux et leur activité de publication instantanée et quasi-inconsciente.

Dans l'état actuel du droit européen, le traitement des données particulières touchant à des informations sensibles comme les opinions politiques nécessite normalement le consentement, avec des exceptions comme la nécessité pour des fins scientifiques, ou les informations déjà rendues publiques par la personne concernée[41]. Le droit prévoit donc déjà une exception basée sur le geste de publication sur laquelle l'approche fondée sur le geste de publication pourrait se baser.

Le geste de publication n'est cependant pas la panacée des problèmes de respect de la vie privée. Un problème important posé par cette approche est le caractère changeant du geste de publication. Comment faire face aux changements de statut d'une information ? Twitter demande aux chercheurs utilisant ses données de supprimer de leurs bases de données tout tweet ayant été supprimé par l'utilisateur, ou étant passé d'un statut public à un statut privé. Cela suppose de mettre en place des dispositifs techniques permettant de suivre le statut des contenus publiés après leur publication. Il existe déjà de tels outils de mise à jour du statut des données, mais il reste à voir quelle est leur portée exacte[42]. Ces problématiques de mise à jour viennent soutenir l'idée que l'approche éthique du développement doit être étendue à l'intégralité du cycle de vie du modèle.

Toutes ces mesures de mise à jour des bases de données perdent beaucoup de leur puissance une fois que la recherche a été publiée, et que les informations ont pu être lues, recopiées et diffusées par des tiers. On retrouve un dilemme lié à l'inéluctabilité de la publication : il est à la fois naturel de penser avoir le droit de retirer une publication et aussi naturel de penser avoir le droit de répéter et disséminer un contenu publié. Une fois un certain degré de dissémination atteint, le retrait d'une publication n'a plus qu'un effet marginal.

Ces problèmes de mise à jour des données sont généralisés par l'introduction dans le RGPD du droit à l'effacement de données, souvent nommé « droit à l'oubli », et du droit à l'opposition au traitement. Le droit à l'oubli est une nouveauté du RGPD : il permet d'exiger l'effacement de données personnelles sans avoir à fournir de justification. Là aussi, il existe une exception pour les activités de recherche mais uniquement si cela compromet le traitement. Le droit d'opposition (au traitement) s'applique en revanche sur la base d'une justification basée sur la situation particulière de l'individu et peut être ignoré si le traitement est nécessaire à une mission de service public (article 21). Tous ces nouveaux droits, comme l'approche fondée sur le geste de publication, confèrent une grande importance à la mise à jour des données, mise à jour qui n'est pas seulement fondée sur la rectification de données inexactes mais sur l'exercice de différents droits du sujet lui permettant de retirer une information de l'espace public.

Cependant, les problèmes posés par une approche fondée sur le geste de publication vont encore plus loin que la mise à jour. Le droit vise à donner un pouvoir de contrôle des individus chaque fois qu'il y a traitement. La republication de données brutes à un autre endroit que celui de leur publication initiale peut être problématique, même si ces données avaient été au préalable publiées à un autre endroit. Le geste de publication ne consiste donc pas un solde de tout compte pour le respect du droit des données. Ce problème contre-intuitif avait été mis en lumière par l'affaire DisInfoLab, où un chercheur faisant face à une controverse sur la qualité et les possibles biais politiques de sa recherche avait publié ses données brutes, essentiellement des publications sur Twitter. On pourrait penser qu'un tel geste ne serait en aucun cas problématique d'un point de vue juridique, même si les tweets en question portent souvent sur des opinions politiques, dans la mesure où ces tweets étaient et demeuraient publics de par la volonté de leur auteur. Ce n'est cependant nullement le cas. L'absence totale de mesure de pseudonymisation ou le choix d'une publication en ligne plutôt qu'une publication au coup par coup à la demande ne sont pas justifiés par des fins de recherche, et sont donc très problématiques car explicitement interdits par le décret

[41] Pour plus de détails, voir le deuxième post du Blog S.I.Lex sur l'affaire DisInfoLab mentionné plus haut.
[42] Pour plus de détails, voir la charte de l'*Association of Internet Researchers (AoIR)* mentionnée plus haut.

d'application de la loi du 1er Août 2018. Le droit français, dans son état actuel, distingue donc le geste de publication d'une permission de republier à volonté.

Mais au-delà de ces problèmes de rectification ou de droit à l'oubli, le geste de publication n'est innocent que s'il porte sur des données affectant exclusivement le sujet de données. Or il est bien connu que les problèmes de respect de la vie privée ne sont pas toujours strictement individuels[43]. La publication d'une donnée personnelle peut affecter d'autres sujets de données que le sujet auteur du geste de publication : la publication de mes données génétiques révèle quelque chose sur les données génétiques de mes parents, la publication de détails de ma vie sexuelle révèle quelque chose sur la vie sexuelle de mes partenaires. Non seulement le geste de publication ne résout pas tout problème de respect de la vie privée, mais il peut être source d'infractions à ce droit.

Ici comme ailleurs dans notre réflexion, l'enjeu exact d'une telle rupture avec la pratique dépend fortement de la forme institutionnelle qu'elle pourrait prendre, selon qu'on souhaite que les modèles présentant un risque pour la vie privée soient prohibés ou qu'on souhaite qu'ils soient signalés. L'absence de données privées dans l'entraînement d'un modèle pourrait devenir un label de qualité affichée par les producteurs, qui pourraient alors livrer leur modèle à l'inspection publique pour démontrer que la norme a été respectée. Cependant, une telle rupture en douceur serait inenvisageable pour les informations les plus sensibles, qui n'ont pas vocation à être diffusées du tout sans l'approbation du sujet de données. Pour éviter tout problème, il devrait donc pouvoir être établi que les informations publiées ne concernent que le sujet de données lui-même. Mais l'approche retomberait alors sur les redoutables problèmes d'opérationnalisation de la notion de référence que nous avons déjà évoqués au sujet de l'exclusion des données personnelles, et perdrait donc un de ses principaux avantages.

Pour résumer, l'approche par exclusion de certains types de données se décline sous différentes modalités, dont toutes affrontent des problèmes pratiques de faisabilité et des problèmes fondamentaux d'opérationnalisation des concepts juridiques. Plutôt que de défendre une modalité particulière comme la « bonne approche », nous préférons présenter ici une suite d'options basées sur des restrictions différentes selon le type de données considéré. Le développeur pourra alors déterminer où il se situe dans cette série d'options, selon les contraintes propres à son projet de développement en termes de faisabilité techniques, d'opérationnalisation des concepts juridiques, du caractère sensible des données considérées et d'autres arguments éthiques et politiques contextuels justifiant le choix d'une option plutôt qu'une autre. La considération de l'ensemble de ces options permet au développeur de s'interroger sur le type de données dont il a vraiment besoin pour accomplir sa tâche, et sur les difficultés techniques posées par la restriction à ce type de données.

La première option, l'exclusion des données personnelles de l'entraînement, a le mérite de proposer une barrière de sécurité dure, qui n'existe pas à l'heure actuelle pour les attaques sur les modèles de ML encodant des données. C'est la raison pour laquelle nous l'incluons dans nos recommandations dans le principe 4. Ce principe a le défaut d'imposer un critère extrêmement strict, excluant des données personnelles dont la diffusion n'est pas problématique, et même parfois désirée : elle pourrait être à la fois indûment contraignante pour les développeurs et rencontrer une résistance des personnes physiques qu'elle cherche à protéger, parce que ceux-ci peuvent désirer la diffusion de certaines informations. Enfin, elle suppose un problème d'opérationnalisation du concept juridique de donnée personnelle. Pour ces raisons, nous proposons une approche graduée de l'exclusion des données personnelles de l'entraînement, qui est exprimé dans la suite de principes 4 à 6. Le principe 4 exprime l'approche la plus exclusive, mais en cas d'inapplicabilité pratique ou

[43] Le droit à la protection des données est à l'heure actuelle un droit purement individuel, qui ne s'applique pas aux groupes. Mais cela n'empêche pas que l'information sur une personne puisse intrinsèquement être une information sur une autre, qui peut aussi faire valoir ses droits sur cette donnée.

d'absence de pertinence d'une exclusion totale pour le respect de la vie privée, nous incluons un renvoi vers les principes 5 et 6 pour des approches plus souples :

4. Entraîner son modèle sans avoir recours à des données personnelles. Si ce principe est inapplicable ou sans pertinence, examinez les options offertes par les principes 5 ou 6 selon le contexte technique et éthique.

Les principes 5 et 6 offrent la possibilité d'une exclusion plus restreinte des données personnelles, qui limiterait la collecte respectivement à des données non-problématiques pour le respect des droits de la personne ou à des données ayant fait l'objet d'un geste de publication explicite, en prenant garde aux problèmes posés par la mise à jour et le retrait des données :

5. Entraîner son modèle sans faire usage de données personnelles dont la diffusion pourrait porter atteinte aux droits des personnes.

6. Entraîner son modèle en faisant exclusivement emploi de données ayant fait l'objet d'un geste explicite de publication. Mettre en place toutes les mesures possibles, à la fois automatiques et manuelles, pour permettre la mise à jour des données, afin de prendre en compte la correction de données erronées, le retrait de publication et l'exercice des droits de rectification, d'effacement et d'opposition.

2. L'approche de sécurité post hoc

Si les modèles entraînés sur des données personnelles sont développés, on peut envisager de prendre toutes les mesures disponibles pour empêcher la rétro-ingénierie des données. À l'approche de respect de la vie privée dès la conception offerte par la restriction des données d'entraînement, on substitue alors une autre approche de *privacy by design* qui consiste à jouer non sur les données, mais sur les propriétés du modèle, afin de limiter les possibilités de rétro-ingénierie des données : nous parlerons donc de respect de la vie privée par défense contre la rétro-ingénierie. Une telle approche souffre cependant d'un manque de définition des boites à outils idoines. Par contraste, une application comme *Exodus Privacy*[44] permet de détecter automatiquement la présence d'outils de traçage parce que le problème du traçage est simple. Le problème d'inversion des modèles est un problème bien plus complexe : la bonne position est bien plus problématique et par conséquent la définition des outils pertinents l'est aussi. Toutes ces difficultés étant bien comprises, les développeurs doivent prendre toutes les mesures possibles pour empêcher la rétro-ingénierie des données, et montrer que le paquet de mesures sélectionnées est à jour de l'état de l'art, obligation énoncée dans le principe 7 :

7. Si le recours à des données personnelles est inévitable, déclarer les raisons justifiant ce recours, ainsi que toutes les mesures prises pour lutter contre la rétro-ingénierie des données, et prendre position sur leur complétude par rapport aux boites à outils et aux méthodes d'attaque existantes.

Il est aussi souhaitable de diffuser les outils de rétro-ingénierie en licence libre, à la fois pour faciliter leur accès et permettre la vérification publique de leur qualité (principe 8) :

8. Diffuser en licence libre tous les outils de sécurisation contre la rétro-ingénierie des données.

2.1 Une autre approche : la certification par la vérification

[44] https://exodus-privacy.eu.org/en/

Afin de démontrer à tous que leur modèle ne permet pas de reconstituer des données personnelles ou d'inférer l'appartenance d'un sujet de données à une base, une approche méthodologique simple serait d'ouvrir le modèle à la vérification. Le modèle pourrait ainsi être mis en ligne et offert aux attaques de tous les hackeurs, qui pourront témoigner que le modèle ne contient aucune faille.

Le premier problème posé par une telle approche est qu'il n'existe pas, comme nous l'avons expliqué dans notre état de l'art, de protection absolue contre la rétro-ingénierie. Dans une telle configuration technique, la vérification des propriétés de sécurité est confrontée à un dilemme : soit on se contente d'une déclaration des mesures de sécurité par le producteur du modèle, ce qui pose les évidents problèmes de délégation et de confiance que la vérification publique était censé résoudre, soit on procède malgré tout à une vérification publique, et on crée une faille de sécurité. Bien que la simple déclaration des propriétés de sécurité sans mise à disposition du modèle pose un problème de délégation, cela permettrait de « se compter » et de forcer les développeurs à sortir du bois, ce qui n'est jamais un effet politique totalement négligeable. Cependant, cela aurait aussi comme effet secondaire, si les techniques de sécurité employées sont *open source*, de faciliter la tâche des attaquants, en leur donnant une liste d'approches à ne pas employer.

Le processus de vérification publique pose donc un problème grave, celui du *vérificateur voleur* : si la personne testant le niveau de sécurité du modèle découvre des données personnelles, qu'est-ce qui l'empêche de s'emparer de ces données et d'en faire un usage malfaisant ?

Il pourrait exister une solution technique au moins partielle à ce souci, qui serait de construire une plateforme sécurisée pour effectuer ces tests, que nous nommerons ici « plateforme de vérification ». Cette plateforme disposerait de tous les moyens de stockage, des moyens de calcul et des bibliothèques de logiciels nécessaires à la tâche de vérification, mais serait sécurisée afin d'empêcher toute fuite de données.

Si une telle institution était techniquement faisable, elle pourrait sembler une solution complète au problème du vérificateur voleur. Cependant, il n'est pas impossible que le vérificateur découvre des données personnelles dans un modèle auquel il aura un accès au moins visuel de par son interface. S'il est dans l'impossibilité d'extraire massivement ces données du modèle par un téléchargement, il peut les voir, en prendre note et les mémoriser, ce qui constitue une potentielle faille de sécurité.

Cette faille de sécurité pose des problèmes d'organisation institutionnelle d'une grande importance. Pour la réduire, l'une des solutions institutionnelles les plus évidentes consiste à n'employer comme vérificateurs que des personnes dûment identifiées, légalement contraintes par un accord de confidentialité voire une assermentation, soumises à une obligation de déclarer toute découverte de données personnelles et à une traçabilité de leurs actions sur la plateforme de vérification. L'accès à la plateforme pourrait être contrainte par des procédures d'identification stricte, comme celles employant des marqueurs biologiques. De telles contraintes permettent de garantir une forte probabilité d'identification du coupable en cas de fuites de données dues à l'interaction visuelle avec les données, et constitue donc un garde-fou légal significatif.

En outre, une telle plateforme, si elle doit être plus qu'une apparence, doit s'assurer de dominer l'état de l'art, et de soumettre un modèle à l'intégralité des attaques possibles : nous appellerons ce problème le problème de la complétude de la vérification. La boite à outils employée pourrait être rendue publique afin que sa pertinence et ses propriétés puissent être discutées publiquement, et chaque déclaration des propriétés de sécurité d'un modèle doit s'accompagner d'une prise de position explicite sur la complétude des mesures prises, et d'une publication des outils de lutte contre la rétro-ingénierie, comme nous l'avons déjà mentionné dans notre discussion des principes 7 et 8.

Si l'on combine les solutions proposées au problème du vérificateur voleur et au problème de la complétude de la vérification, on se retrouve avec un dispositif ressemblant plus à une autorité

de contrôle industriel[45] qu'à un dispositif classique de la culture du libre, qui serait la mise en ligne gratuite du modèle, à disposition des hackers du monde entier. Les enjeux de sécurité posés par la récupération des données personnelles à partir du modèle semble interdire de traiter les modèles comme des objets classiques du libre, du moins tant que ces problèmes de sécurité n'ont pas été résolus. Une fois un modèle déclaré vide de toute donnée personnelle, ou imperméable à la rétro-ingénierie de ces données, celui-ci pourrait parfaitement faire l'objet d'une diffusion gratuite sous licence libre.

Le dispositif de plateforme de vérification que nous venons d'esquisser est donc étranger à la culture décentralisée du libre, souvent marquée par des idéaux libertaires[46]. Les enjeux de sécurité poussent à la formation d'un dispositif sécurisé, contrôlé et mis à jour par un personnel dédié, et qui pourrait devenir à terme un organisme de certification. Dans un tel dispositif, on retrouvera évidemment les problèmes de confiance, et de *quis custodiet custodies ?,* que pose toute délégation d'une tâche sensible à une organisation centralisé de contrôle dont les activités légitiment des restrictions d'accès à et de diffusion de l'information. Cependant, il serait possible de faire suivre une vérification sécurisée par une vérification publique : les vérificateurs mettraient en libre accès tous leurs moyens sur une plateforme ouverte, idéalement avec des bibliothèques logicielles en *open source*, afin que chacun puisse vérifier par lui-même que les vérifications opérées sont honnêtes. La seule restriction pourrait venir des modèles propriétaires, qui ne pourrait être soumis qu'à des vérifications en boite noire, mais les modèles *open source* pourraient être soumis à un processus entièrement transparent de re-vérification. Une telle approche n'est néanmoins possible sans créer de failles de sécurité, comme nous l'avons dit plus haut, que si le résultat final de la vérification sécurisée est qu'il n'existe aucun moyen d'extraire des données personnelles du modèle. On se retrouve donc face à un dilemme politique classique : en l'absence d'une sécurisation parfaite, soit l'on choisit un geste de publication constituant une faille de sécurité, soit on adopte des procédures de restriction d'accès à l'information posant de graves problèmes de délégation et de centralisation du pouvoir.

La résolution de ces problèmes fondamentaux de politique institutionnelle irait bien au-delà de l'ambition de ce document. Nous nous contenterons ici d'encourager les développeurs, lorsque les failles de sécurité créées sont tolérables, à ouvrir leur modèle à la vérification, en détaillant les raisons de la décision, les mesures prises pour mitiger le problème du vérificateur voleur et en indiquant leurs positions par rapport à l'état de l'art. Il est aussi souhaitable que ceux qui choisissent de ne pas offrir leurs modèles à la vérification publique offrent leurs raisons, afin que la communauté puisse avoir un débat informé sur les différents types d'arbitrage entre sécurité, circulation de l'information scientifique et construction de la confiance par la transparence (principe 9) :

9. Si cela n'entraîne pas de faille de sécurité intolérable, mettre le modèle à disposition de tous afin de permettre la vérification publique des propriétés de sécurité. Détailler les raisons de la décision positive ou négative, les mesures de sécurité prises contre les failles créées par cette mise à disposition du public, et prendre position sur leur complétude par rapport à l'état de l'art.

[45] Il est naturel de songer ici à une agence gouvernementale capable d'offrir une certification reconnue par la loi. Rien n'interdit cependant que cet organisme de contrôle soit une fondation indépendante de l'État. Ceci poserait évidemment la question du financement de cette plateforme, qui devrait garantir la suffisance et la pérennité des moyens tout comme l'indépendance de l'organisme.

[46] On retrouve une expression typique de ce courant de pensée dans l'*Ethical Design Manifesto*, qui affirme clairement le privilège donnée aux infrastructures décentralisées pour des raisons proprement morales et politiques : « *Technology that respects human rights is decentralised, peer-to-peer, zero-knowledge, end-to-end encrypted, free and open source, interoperable, accessible, and sustainable.* » « Ind.ie — Ethical Design Manifesto », consulté le 6 septembre 2020, https://2017.ind.ie/ethical-design/.

2.2. L'acceptation de la restriction d'accès pour des cas exceptionnels

La culture de diffusion libre des connaissances logicielles que nous promouvons ici n'est pas nécessairement exclusive d'un usage restreint d'une approche en boite noire. La culture libriste est avant tout une approche par défaut, qui défend que le logiciel devrait être en règle générale conçu comme un bien scientifique commun : elle peut admettre des exceptions. Prenons par exemple les modèles d'apprentissage tâchant de détecter l'apparition et la propagation d'une épidémie par collecte et analyse en temps réel des propos tenus sur les réseaux sociaux[47]. Un tel modèle est évidemment en rupture avec l'approche de respect de la vie privée dès la conception par restriction des données utilisées que nous avons présentée plus haut. Il reste à voir si son développement en temps réel est compatible avec une approche par défense contre la rétro-ingénierie. En outre, même la vérification publique du modèle pourrait être problématique dans ce contexte : elle pourrait prendre trop de temps pour un dispositif à déployer dans l'urgence, ou être confronté à des instances trop graves du problème du vérificateur voleur.

Nous pourrions donc être confrontés à des cas de figure où toutes nos mesures de respect de la vie privée -restrictions diverses de la collecte, lutte contre la rétro-ingénierie, et vérification publique- seraient inopérantes, mais où le modèle se révélerait d'une grande utilité pour prévenir et lutter contre les épidémies. On pourrait alors envisager une exception à la règle générale de publicité et aux mesures de respect de la vie privée dès la conception envisagées ici. La vie privée serait alors protégée par des restrictions d'accès sévères au modèle comme à ses bases de données d'entraînement et de test, et par un cadre législatif serré prohibant tout usage du modèle ou des bases de données pour des fins autres que celles ayant justifié l'exception à la règle, et éventuellement une destruction des bases de données à la fin de la période d'urgence. Nous énonçons cette exception à notre approche générale dans le principe 10 :

10. *Lorsque les mesures de restrictions de la collecte et de lutte contre la rétro-ingénierie ne sont pas applicables, et que la gravité de l'enjeu surpasse les enjeux de vie privée, autoriser un modèle encodant des données privées en restreignant strictement l'accès à ce modèle et son emploi pour l'usage ayant justifié l'exception. La discussion de l'exception faites au respect de la vie privée et à la publicité de la connaissance scientifique devra prendre en compte les propriétés singulières des modèles de ML, comme la capacité à apprendre en temps réel sur une grande masse de données, l'opacité du fonctionnement et de son évolution, et la qualité de leurs prédictions comparée à d'autres modèles.*

On sait les faiblesses et les dangers d'une telle approche : la mise à disposition d'un outil puissant et invasif crée dans les institutions une tentation, bien souvent irrésistible, d'en élargir progressivement ou soudainement l'usage à des cas de plus en plus nombreux, et de plus en plus favorables au contrôle social. La crainte de la banalisation des mesures d'exception, si légitime soit-elle, ne clôt cependant pas le débat sur leur nécessité : il n'est pas facile de refuser un outil qui pourrait sauver des vies. D'un point de vue éthique et politique, il reste donc à décider ce qui doit primer, de la volonté de faire tout son possible dans une situation d'urgence et d'exception, ou de lutte préventive contre les dérives institutionnelles à long terme si souvent enclenchées par ces situations. Mais l'objet de ce texte n'est pas de prendre position sur ce débat classique de philosophie politique et de stratégie institutionnelle. Il est plutôt de prendre acte de la présence de ce problème, et de tâcher de voir si les modèles d'apprentissage automatique soulèvent des enjeux singuliers, ou ne mènent qu'à la reproduction d'un espace de positions philosophiques préexistant.

Si nombre d'aspects de ce débat philosophique seront assurément inchangés, on peut distinguer une singularité des modèles de ML digne d'être commentée. L'avantage crucial de l'apprentissage automatique pour ce type d'usages est qu'il permet un ajustement du modèle en temps réel à un flux massif de données. Il s'agit là d'un avantage tactique décisif pour faire face à des situations d'urgence dont les évolutions rapides et imprévues peuvent défaire les suppositions d'un modèle classique conçu en amont. Un tel avantage suppose cependant une collecte massive

[47] Aditya Joshi et al., « Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective », *ACM Computing Surveys (CSUR)* 52, nº 6 (2019): 1–19.

de données en temps réel, et peut entraîner des évolutions imprévues des capacités du modèle. Ces évolutions pourraient comprendre le développement d'une capacité prédictive trop fine, ou une plus grande sensibilité à des failles de sécurité comme la rétro-ingénierie. La collecte massive de données en temps réel et l'évolution dynamique du modèle présentent toutes deux des risques majeurs pour le respect de la vie privée. L'emploi des modèles de ML pour faire face à des situations d'urgence renforce à la fois des arguments *pro* -possibilité d'une capacité prédictive accrue et d'adaptabilité du modèle en temps réel- et *contra* -collecte massive et rapide de données, manque de visibilité sur les évolutions du modèle. La discussion d'une possible exception à la publicité des modèles et au respect de la vie privée dès la conception se devra donc de prendre en compte ces propriétés particulières des modèles de ML. Nous incluons donc cette prise en compte dans la deuxième phrase de notre principe 10.

3. Le désentraînement des modèles

Il s'agit d'une approche sans résultats systématiques, qui consisterait à modifier un modèle entrainé sans reproduire l'intégralité de l'entraînement, mais en empêchant certaines inférences préalablement possibles comme des inférences d'appartenance par exemple. On peut par exemple filtrer a posteriori les résultats avant de les afficher. De tels filtres pourraient être utiles pour résoudre les problèmes imprévus, pratiquement inévitables, qui se présenteront un par un. Mais en vie privée plus encore qu'ailleurs, il vaut mieux prévenir que guérir, puisque la signalisation d'un problème peut souvent signifier que le mal a déjà été fait. L'existence de tels filtres n'est évidemment pertinente que si l'accès au modèle lui-même est hautement sécurisé.

Cette approche devrait assurément faire l'objet d'efforts notables dans les années à venir, tant elle pourrait offrir des solutions adaptées aux nouveaux problèmes posés par les capacités d'inférence statistiques offertes par le ML. Elle serait particulièrement intéressante pour une approche de la protection des données personnelles, comme celle du « droit à l'inférence raisonnable » de Wachter & Mittelstadt[48] que nous présenterons dans la section III, qui se base non plus sur une restriction de l'accès aux données mais sur une restriction du pouvoir d'inférence des modèles.

4. La confidentialité différentielle

La confidentialité différentielle (*differential privacy*) est fondée sur une certaine conception théorique du respect de la vie privée par un algorithme[49]. Un algorithme est confidentiel si l'analyse des données qu'il produit ne révèle pas beaucoup plus sur un sujet de données que si l'analyse était effectuée sur un ensemble de données où il ne figurerait pas. Elle consiste à ajouter du bruit aux requêtes pour que la suppression de la mention d'un sujet de données ne modifie pas sensiblement le résultat des requêtes. L'information accessible est donc réduite à celle qui est partagée avec les autres sujets de données, par opposition aux informations spécifiques à un sujet précis.

La définition de la confidentialité différentielle est faite pour des algorithmes randomisés. Pour un algorithme déterministe, si on dispose de deux ensembles de données voisins, c'est-à-dire ne différant que par une entrée, il est aisé d'inférer une information sur cette entrée si les deux ensembles ont une sortie différente pour une même requête. Par exemple, si un algorithme déterministe donne la moyenne des revenus des individus figurant dans une base de données, il est aisé d'inférer le revenu d'un individu par l'impact du retrait de ses données sur la moyenne globale. Intuitivement, un algorithme randomisé respectant la confidentialité différentielle donnera en sortie des distributions de probabilités très similaires pour deux ensembles de données voisins. Pour les techniciens, on peut dire que la confidentialité différentielle (ε, δ)-DP assure que la valeur absolue de la perte de confidentialité (*privacy loss*) est bornée par ε avec une probabilité (1-δ). La quantité ε

[48] Sandra Wachter et Brent Mittelstadt, « A right to reasonable inferences: re-thinking data protection law in the age of big data and AI », *Colum. Bus. L. Rev.*, 2019, 494.

[49] Cynthia Dwork et al., « Our data, ourselves: Privacy via distributed noise generation », in *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, 2006), 486–503.

est appelée « budget de confidentialité » (*privacy budget*) : chaque calcul, requête ou tabulation consomme une fraction de ce budget. L'approche permet donc de quantifier le risque de perte de confidentialité, et d'effectuer des choix tactiques en affectant une plus grande part du budget de confidentialité à des tables et calculs considérés plus essentiels. L'approche est robuste contre toute connaissance auxiliaire. Elle permet aussi de publier l'algorithme, et a l'avantage de résister aux attaques par analyse de la sortie (*postprocessing*) et d'être compositionnelle. Enfin, si nous la présentons ici parmi les approches post hoc, la confidentialité différentielle pourrait tout aussi bien être classée parmi les approches vertueuses dès la conception, puisque nombre d'implémentations interviennent dès la phase d'apprentissage (pour un cours d'introduction, voir[50]).

Puisqu'un modèle est censé apprendre des traits généralisables et non régurgiter par cœur des points de données, une telle approche ne devrait pas en théorie être dommageable au développement des modèles. Reste dans la pratique à effectuer un arbitrage subtil entre confidentialité différentielle et utilité, la méthode privilégiée actuelle d'ajout de bruit pouvant bien sûr avoir un impact sensible, voire dramatique dans certains cas, sur les performances du modèle. En plus de disposer d'un certain prestige théorique, l'approche de confidentialité différentielle est en plein essor dans les applications industrielles, et est adoptée par certaines grandes entreprises et certaines institutions publiques comme le US Census Bureau[51].

La confidentialité différentielle est aussi compatible avec une approche d'apprentissage décentralisé ou fédéré : au lieu de centraliser les données d'apprentissage chez un tiers de confiance, les données sont conservées par le sujet de données. La confidentialité différentielle offre alors la possibilité aux sujets de bruiter leurs données à la source. Le coût en utilité pour une même garantie de confidentialité par rapport à une approche centralisée peut cependant être important. Il existe des stratégies de contournement de ce problème, mais le coût computationnel peut être élevé, limitant l'approche à certains calculs ou à un certain nombre de participants[52],[53],[54].

D'un point de vue plus fondamental, l'approche est plus facile à appliquer lorsque tous les individus ont le même poids : les sujets correspondant à des points de données exceptionnels, ou *outliers*, sont plus difficiles à protéger[55]. Les *outliers* peuvent être des individus ayant un grand besoin de sécurité de par les propriétés qui font d'eux des *outliers*, que ce soit par la fortune personnelle, l'exposition médiatique ou le nombre de menaces de mort reçues. La question de la protection des *outliers* est d'autant plus importante pour la confidentialité différentielle que celle-ci a été conçue en opposition aux approches dites « *just a few* », qui admettent la possibilité de sacrifier quelques individus pour protéger le groupe, et vise à la protection de chaque sujet de données. La protection des informations sur les *outliers* est structurellement difficile, puisqu'elle peut nécessiter une telle quantité de bruit que les réponses aux requêtes deviennent absurdes. Il existe cependant des stratégies visant à maîtriser ce problème, notamment en cherchant à créer des modèles statistiques moins sensibles à la présence d'un individu[56]. D'autre part, réduire l'information accessible sur un individu à l'information partagée avec d'autres sujets de données n'est pas la panacée du respect de

---

[50] Aurélien Bellet, « Privacy Preserving Machine Learning », consulté le 15 décembre 2020, http://researchers.lille.inria.fr/abellet/teaching/private_ML_course.html.

[51] Michael Hawes, « Title 13, Differential Privacy, and the 2020 Decennial Census », s. d., 32.

[52] K. Wei et al., « Federated Learning With Differential Privacy: Algorithms and Performance Analysis », *IEEE Transactions on Information Forensics and Security* 15 (2020): 3454-69, https://doi.org/10.1109/TIFS.2020.2988575.

[53] Qiang Yang et al., « Federated Machine Learning: Concept and Applications », *ACM Transactions on Intelligent Systems and Technology* 10, nº 2 (28 février 2019): 1-19, https://doi.org/10.1145/3298981.

[54] T. Li et al., « Federated Learning: Challenges, Methods, and Future Directions », *IEEE Signal Processing Magazine* 37, nº 3 (mai 2020): 50-60, https://doi.org/10.1109/MSP.2020.2975749.

[55] Voir les références discutées dans Michael Veale, Reuben Binns, et Lilian Edwards, « Algorithms that remember: model inversion attacks and data protection law », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, nº 2133 (2018): 20180083.

[56] Edward Lui et Rafael Pass, « Outlier privacy », in *Theory of Cryptography Conference* (Springer, 2015), 277–305.

la vie privée. Comme nous l'avons déjà vu, la simple appartenance à une base de données -d'anciens détenus, de patients, de membres d'un parti politique, …- constitue en soi une information sensible, protégée par de nombreux cadres juridiques.

La confidentialité différentielle est une des grandes approches théoriques du respect de la vie privée pertinente pour les modèles d'IA, et ses applications industrielles sont en plein développement. Son emploi mérite donc d'être considéré par les informaticiennes et informaticiens entraînant des modèles sur des bases de données personnelles. Elle permet en théorie non seulement de prévenir des attaques en boite noire, mais aussi, dans ses versions intégrées à l'apprentissage, de prévenir des attaques en boite blanche comme la rétro-ingénierie des modèles, ou des attaques fondées sur le pouvoir prédictif excessif ou la suroptimisation : en ajoutant du bruit aux données dès l'apprentissage, on empêche l'apparition de l'ensemble des problèmes dus à un apprentissage trop fin. L'ensemble des protections offertes est cependant probabiliste par nature, et ne constitue pas une barrière de sécurité absolue : reste à voir si dans les cas d'usage concrets, des protections approchant la quasi-certitude pourront être atteintes sans réduire l'utilité du modèle à néant. Cette charte n'est pas consacrée à la sécurité au sens technique, et nous ne ferons donc pas du respect de la confidentialité différentielle un principe. Il est certain que si cette approche était généralisable avec succès à nombre d'applications industrielles, l'importance de beaucoup de problèmes évoqués dans cette charte en serait grandement diminuée, ce dont on ne pourrait que se réjouir.

III. Conséquences philosophiques et juridiques de l'état de l'art

Le brouillage de la distinction entre données et logiciel

Une profonde conséquence juridique des problèmes de sécurité des modèles est le possible brouillage des catégories juridiques censées protéger les données personnelles. Comme l'ont remarqué Wachter et Mittelstadt dans *The right to reasonable inference* et M. Veale et al. dans *Algorithms that remember*, les données brutes et les données traitées par un algorithme sont soumis à deux régimes juridiques différents dans la législation européenne. Les premières font l'objet d'un droit dédié, le droit de protection des données, qui les munit de nombreuses garanties légales. Les données traitées, ne serait-ce que légèrement, par un algorithme passent sous un régime juridique différent, où la protection des données personnelles ne fait plus l'objet d'un droit dédié, et où les droits de la propriété intellectuelle et du secret commercial s'exercent avec bien plus de puissance. Ce modèle juridique posait déjà problème avant les modèles d'apprentissage automatique, dans la mesure où nombre de traitements des données étaient déjà invertibles, et permettaient donc de récupérer les données personnelles initiales. Mais les modèles d'apprentissage automatique ont ceci de particulier que les données sont encodées dans ce qui compte à l'heure actuelle comme un logiciel, et sont donc soumises à un régime juridique différent. Les modèles d'apprentissage automatique viennent donc brouiller la distinction actuelle entre données d'une part, et algorithmes, programmes et codes source d'autre part, ce qui pose des enjeux considérables de redéfinition de catégories juridiques fondamentales.

Dans *Algorithms that remember*, Veale et al. argumentent que les modèles susceptibles d'une attaque par inversion ou par inférence d'appartenance pourraient être considérés comme les données pseudonymisées ou soumises à un autre traitement cryptographique. Une telle analogie vient faire exploser la barrière entre données traitées et données brutes, mais elle n'est pas complètement étrangère à l'état actuel du droit. Les données pseudonymisées ont bien fait l'objet d'un traitement, qui plus est d'un traitement explicitement conçu pour protéger les droits des sujets de données, mais elles sont soumises au même statut juridique que les données personnelles dont elles sont tirées, et constituent déjà une exception au cadre juridique général des données traitées.

Ce point important mérite d'être développé plus avant. La capacité d'inférence de l'identité d'un sujet à partir de données anonymisées ou pseudonymisées est déjà prise en compte par le droit

européen, ce qui veut dire que le problème d'inversion est aussi déjà pris en compte. La probabilité de ré-identification est même conseillée comme mesure pour déterminer si un ensemble de données est personnel ou non dans les dernières interprétations du RGPD (voir plus bas), ce qui fait même de la possibilité d'inversion une partie de la définition du caractère personnel des données. De plus, une donnée peut être considérée comme personnelle même si l'ensemble de données servant à la désanonymiser est entre les mains d'une autre entité. Enfin, on peut évaluer qu'une information réfère à une personne physique par contenu, mais aussi par objectif ou par résultat (voir section *L'état de l'art juridique sur le concept de donnée personnelle*). La catégorie de données personnelles est donc loin d'être simple, il n'est même pas sûr qu'elle soit intrinsèque à un ensemble de données si on inclut l'objectif. Avec une telle définition large, même une information d'appartenance pourrait être considérée comme une information personnelle. Ceci montre qu'il existe déjà dans le droit actuel des ressources pour faire face aux problèmes posés par l'inversion des modèles, et qu'il serait donc possible d'étendre les définitions actuelles pour inclure les modèles d'apprentissage automatique s'ils sont susceptibles d'une attaque par inversion permettant de retrouver des données personnelles.

Il faut noter que cette définition par capacité de récupération ne s'accorde que très peu avec l'intuition que la donnée transformée par le traitement algorithmique ne peut plus être considérée comme une donnée, qui est aussi présente dans le droit européen. Une telle conception semble fondée sur une légitimation de la propriété privée par le travail et l'investissement financier. Lorsqu'on considère que le sujet de données est propriétaire de ces données personnelles, on s'appuie sur une philosophie de la propriété bien différente, et éventuellement distincte du concept de propriété privée tel qu'il est compris par le droit, car dépourvue de toute relation avec le travail et l'investissement, mais fondée sur le rapport à soi. Lorsque le droit des données personnelles se voit donner prééminence sur la propriété intellectuelle ou le secret des affaires, on retrouve l'intuition que la vie privée constitue un droit fondamental qui écrase le droit de propriété, ou qui constitue un droit de propriété privilégié : je suis le propriétaire de mes données, même si c'est vous qui travaillez avec et sur elles.

La prise en compte des enjeux de vie privée posés par les modèles de ML appelle donc à une reconsidération radicale du droit européen des données. Les distinctions juridiques entre données brutes et données traitées, données et logiciels doivent être repensées si l'on veut que le droit demeure à la fois cohérent et adapté aux évolutions technologiques. Cette nouvelle conception du droit des données devrait être fondée sur une meilleure articulation philosophique entre le droit de propriété fondé sur le travail, et le droit de propriété fondé sur le rapport à soi.

Nous allons à présent opérer quelques rappels sur l'état de l'art juridique européen, avant de présenter quelques perspectives sur le futur du droit des données qui sont pertinentes pour nos problèmes.

5. L'état de l'art juridique sur le concept de donnée personnelle

Nous avons vu plus haut (voir section II B.1, *L'exclusion des données personnelles*…) que d'après le Considérant 26 du RGPD[57] les données anonymes sont les données dépourvues de référence à une personne physique, ou traitées de telle manière que les personnes physiques ne soient plus identifiables[58]. Une large partie des difficultés liées à cette définition provient de la subtilité, et de la souplesse, introduite par la modalité : les évolutions technologiques récentes ayant

[57]Ce texte n'est pas contraignant pour les États de l'UE, mais son occurrence dans le texte cardinal de la législation européenne lui donne assurément une autorité certaine dans le débat juridique.

[58]L'une des grandes difficultés de l'anonymisation est qu'une combinaison d'informations anonymes peut permettre la ré-identification d'une personne physique. Voir Aleksandra Drozdz, *ibid.,* note 34 et les références ci-mentionnées. Les données pseudonymisées sont précisément définies dans le RGPD comme les données qui ne peuvent être rapportées à un individu sans l'ajout d'information additionnelle.

considérablement modifié, et continuant à modifier, les conditions sous lesquelles une personne peut être identifiable.

Les données pseudonymisées sont des données personnelles d'après cette définition, puisque le processus de pseudonymisation est invertible. Le considérant 26 du RGPD utilise la « probabilité raisonnable » (*reasonable likelihood*) de ré-identification comme un délimiteur de ce qui constitue une donnée personnelle *qua* donnée ré-identifiable. La probabilité raisonnable est définie en termes « objectifs », c'est-à-dire en fonction des possibilités techniques.

On assiste ainsi à une forme de *datafication* discrète du concept de donnée personnelle. Celui-ci est tout d'abord définie de manière générale comme une information, sans aucune mention d'un éventuel encodage numérique ou d'un éventuel traitement informatique : l'information personnelle pourrait donc prendre toute forme analogique, comme une allusion dans une conversation orale non-enregistrée ou une représentation picturale, comme il est explicitement admis par l'Article 29 Working Party[59]. Celle-ci inclut explicitement tout type de médium pour l'information personnelle. Mais la limite de l'identifiable est définie purement en termes techniques, c'est-à-dire en termes informatiques. Un tel angle d'attaque n'a rien d'innocent, dans la mesure où les capacités de ré-identification humaine ne sont pas les mêmes que celles de ré-identification technologique, et que les premières ne sont pas encore strictement incluses dans les secondes. Comme nous l'avons déjà mentionné plus haut (section II B.1), une allusion contextuelle telle que « tout cela concerne une personne que tu connais bien » pourra être parfaitement transparente pour un être humain, alors qu'elle pourrait être indéchiffrable par un système informatique. Une telle approche technocentrée porte en elle un risque de négligence des capacités d'identification contextuelle des êtres humains, qui sont pourtant cruciales pour le droit. En tout état de cause, cette définition ouvre volontairement la porte à une dépendance contextuelle du statut de donnée personnelle à l'évolution de l'état de l'art technologique.

Comme nous l'avons déjà mentionné, la notion d'identification directe employée par le Considérant 26 est définie comme l'identification par le nom propre, éventuellement secondée par une information distinguant les homonymes, tandis que l'identification indirecte est définie par une combinaison unique d'identifiants permettant de singulariser l'individu au sein d'un groupe. La dépendance au contexte technologique mouvant explique l'opinion complémentaire du *Working Party 29* sur l'anonymisation : une donnée n'est anonyme que lorsque l'anonymisation est irréversible, c'est-à-dire quand il n'est pas possible de retrouver l'identité d'une personne physique à partir de cette donnée avec les moyens existants de la technologie. Cette possibilité de ré-identification peut en outre être comprise dans un sens technique absolu, ou non dans un sens relatif aux moyens à la disposition du contrôleur de données. L'objectif explicite ou implicite du traitement de l'information est employé pour juger de la probabilité raisonnable de ré-identification, car si le traitement n'a de sens que si des personnes physiques sont identifiées, la présence d'outils de ré-identification doit être considérée comme raisonnable. Une information peut concerner une

[59] L'Article 29 Working Party est une autorité de conseil de l'UE ayant émis en 2007 une opinion sur le concept de donnée personnelle pour faciliter le travail des États membres, la WP 136 (*Opinion 4/2007 On the concept of personal data*). Même si ses prises de position n'ont pas valeur contraignante pour les États membres, il semble que sa prise de position sur les données personnelles ait déjà acquis une influence considérable sur la jurisprudence européenne, y compris sur une jurisprudence aussi décisive que celle de la Cour de Justice Européenne : *« While the A29WP's Opinions are not binding, they do carry a large degree of significance in EU data protection doctrine and practice. The Court of Justice of the European Union (CJEU), the chief judicial authority of the EU, tasked with ensuring uniform interpretation of EU law, itself implicitly adhered to the construction of the "relational" set forth by the A29WP's Opinion on the concept of personal data in its YS and Nowak judgements, and explicitly referred to another A29WP opinion in the recent Jehovan todistajat case. National courts and supervisory authorities consider them in their proceedings, too. The A29WP's interpretation of the wording 'relating to', and of the notion of personal data tout court, is thus prominent in the European data protection milieu, cemented by the CJEU's interpretation in the YS and Nowak cases.*" Dalla Corte, *ibid.*

personne physique[60] par son contenu, son objectif ou son résultat. Une information concerne un individu par son contenu lorsqu'elle est fondamentalement une information à propos de cet individu. Une information concerne une personne par son objectif lorsqu'elle est utilisée ou pourra être probablement utilisée pour influencer le statut ou le comportement d'une personne physique, et dans son résultat lorsqu'il est probable qu'elle ait un impact sur une telle personne physique. L'impact n'a pas à être majeur, et il suffit que la personne puisse être traitée de manière différente des autres à partir de cette information. L'information concerne une personne par son résultat si son usage aura probablement un impact sur les droits et intérêts de cette personne, en prenant en compte toutes les circonstances du cas considéré. Là encore, cet impact n'a pas à être majeur, et il suffit que la personne puisse être traitée différemment de par cet impact.

Ce sont évidemment les conditions d'objectif et de résultat qui donnent une portée potentiellement très ample au concept de donnée personnelle. Cette vaste extension de la notion peut être justifiée par le fait que les données dont le contenu fait directement référence à un individu ne sont pas les seules à pouvoir causer un tort informationnel. Toute donnée pouvant être utilisée pour générer des conséquences pertinentes pour le sujet de données peut créer du tort informationnel, et donc tomber dans la portée de l'esprit du droit des données personnelles.

Dans la mesure où une très large partie des données sont aujourd'hui collectées à des fins d'influence sur le comportement des individus, et pourront avoir des impacts multiples sur les personnes physiques à travers les objets connectés, l'Internet et les *smart cities*, elles peuvent devenir personnelles au titre de cette compréhension. C'est précisément le cœur d'un argument remettant en cause la notion de donnée personnelle que nous allons à présent discuter.

La remise en cause de la notion de donnée personnelle

Il existe dans la littérature juridique des approches radicales centrées non pas tant sur un élargissement de la notion de donnée personnelle que sur son dépassement. Les deux exemples que nous allons présenter peuvent être vus comme des réponses juridiques à la « mort de l'anonymat » provoqué par la datafication ubiquitaire et les nouveaux pouvoirs d'inférence statistique[61]. La première, due à Nadezdha Purtova[62], argumente que la définition actuelle, si l'on inclut les positions du Considérant 26 et du WP 136 est absurdement maximaliste[63]. À titre d'exemple provocant, Purtova considère le cas d'une expérience de *smart city* conduite dans la ville d'Eindhoven en Hollande. Une rue dont la fréquentation déclinante aurait été causée par la délinquance et le vandalisme a été massivement équipée de senseurs et de caméras vidéo, dont les données sont massivement analysées afin de prévenir les activités « déviantes ». Dans cette expérience à grande échelle de surveillance, non seulement une large partie des données permettent de ré-identifier des personnes physiques, mais toutes les données sont collectées afin d'influer sur le comportement des personnes passant dans la rue. À ce titre, elles sont toutes potentiellement des données personnelles par objectif et/ou par résultat, y compris les données collectées sur les conditions climatiques : toutes les données collectées sur la rue sont donc devenues des données personnelles d'après le raisonnement de Purtova. Toujours d'après elle, la jurisprudence n'a pas apporté de restrictions substantielles à cette interprétation très large. L'interprétation maximaliste de la notion de données personnelles est donc vouée à devenir un problème de plus en plus aigu à

[60] Il faut remarquer que le droit européen considère que les personnes physiques sont des êtres humains vivants, et que par conséquent les données concernant une personne décédée ne sont pas des données personnelles. Voir Aleksandra Drozdz, *ibid.*

[61] Sur les risques de désanonymisation, et la possible mort technique de l'anonymat, voir par exemple Paul Ohm and Scot Peppet,'What If Everything Reveals Everything?' in Cassidy R Sugimoto, Hamid R Ekbia and Michael Matoli (eds), Big Data Is Not a Monolith (MIT Press 2016).

[62] Nadezhda Purtova, « Purtova: The law of everything. Broad concept of personal data and future of EU data protection law », *Law, Innovation and Technology* 10, nº 1 (2018): 40−81.

[63] Une position similaire est défendue dans Adekemi Omotubora et Subhajit Basu, « Next generation privacy », *Information & Communications Technology Law* 29, nº 2 (2020): 151–173.

mesure que se poursuit le transfert d'activités sur Internet, l'usage grandissant des objets connectés et l'émergence des *smart cities*. Avec la croissance concomitante des exploitations de ces données afin d'inférer des connaissances sur les personnes physiques, le nombre de données pouvant être considérées comme « données personnelles par objectif ou par résultat » pourrait croître indéfiniment, et comprendre pratiquement n'importe quelle information sur l'environnement des individus.

Les évolutions de la technologie ont donc introduit une profonde liquidité de la notion de « donnée personnelle ». La notion est basée sur la capacité d'identifier une personne à partir de la donnée considérée, et c'est précisément cette capacité à identifier qui a été profondément affecté par les évolutions techniques récentes, y compris, et surtout, les techniques venues de l'IA. Cette liquidité créé non seulement une considérable incertitude sur le futur du statut d'une donnée quelconque : elle ouvre la porte à une véritable explosion de l'application de la notion. Dans un nouveau monde où la quasi-intégralité des informations peuvent être datafiées, et où la capacité à ré-identifier à partir de données ne cesse de s'accroître, les lois de protection des données personnelles risquent bel et bien de devenir, pour reprendre l'expression de Purtova, la « loi de tout et n'importe quoi. » Le système juridique de protection des données personnelles est ainsi exposé à un véritable risque de saturation opérationnelle, rendant son application impossible. Pour protéger les droits des personnes physiques, y compris le droit à la vie privée, il serait donc nécessaire de changer de paradigme juridique en dépassant la notion de « donnée personnelle », et en réformant le cadre juridique actuellement centré sur la protection de ces données.

Comme on peut s'y attendre, il existe une variété de propositions sur la direction à prendre à partir de ce constat de faiblesse de l'interprétation maximaliste. Purtova propose ainsi d'aller au-delà de la simple identifiabilité pour parler de la relation de l'information à une personne. Elle propose aussi de dépasser la distinction binaire entre identifiable et non-identifiable, pour donner lieu à une classification intensive. Omotubora & Adekemi défendent que, bien que le concept de vie privée soit centré sur des dommages subjectifs, comme la perte de dignité ou l'atteinte à la réputation, la législation sur les données devrait être centrée sur les dommages objectifs, comme le vol d'identité ou la fraude.

Comme nous l'avons vu plus haut, et comme il a déjà été remarqué par Veale & al. Dans *Algorithms that remember*, la thèse que les modèles de ML sujets à des attaques pourraient être conçus comme des données personnelles ne dépend pas forcément de l'interprétation maximaliste. Elle peut simplement s'appuyer sur une extension de la classification des données personnelles prenant en compte la capacité d'inversion. En outre, et surtout, l'argument de Purtova est basé sur la classification fondée sur l'objectif ou le résultat : il est douteux qu'il passe aux données personnelles par contenu. Au lieu d'abandonner complètement la notion de donnée personnelle, il pourrait être envisagé d'abandonner l'interprétation maximaliste de cette notion, en contraignant fortement la classification par objectif ou par résultat. Dalla Corte défend également une position de ce type.

Ce n'est pas le lieu que de résoudre ce difficile problème juridique. Mais il nous faut noter que sa solution est vouée à avoir un puissant impact sur certaines de nos stratégies de développement éthique. La stratégie consistant à éviter l'emploi de données personnelles dans l'entraînement deviendrait complètement vaine dans un monde où l'interprétation maximaliste de la notion triompherait, et pratiquement tout et n'importe quoi serait une donnée personnelle. Le futur de cette approche dépend donc du futur du débat sur l'interprétation maximaliste. Il dépend en particulier d'une possible réinterprétation plus restrictive des données personnelles, qui éviterait l'interprétation maximaliste en recentrant la notion sur la référence par contenu. S'il s'agit assurément d'un problème difficile pour le Traitement Automatique de la Langue, la reconnaissance d'une référence à une personne physique est une tâche plus restreinte que la reconnaissance d'un objectif ou d'un résultat. Quel que soit le futur du débat juridique sur ces

questions, il est évident que toutes les tentatives d'opérationnalisation des concepts nécessaires au respect de la vie privée dépendent crucialement de la définition des données personnelles.

Le droit à l'inférence raisonnable

Les critiques de la notion de donnée personnelle ne sont pas toutes basées sur cette combinaison d'interprétation juridique maximaliste et de datafication ubiquitaire. Parmi les pistes explorées dans le cadre de cette réflexion, on doit évoquer celle envisagée par Wachter & Mittelstadt dans *A right to reasonable inference*. Prenant acte du changement de paradigme dans la protection des données personnelles, ils défendent que le droit devrait lui aussi changer de paradigme pour se concentrer non plus tant sur les données personnelles et leur accès, mais sur les inférences faites à partir de ces données. Dans la mesure où il est possible d'inférer des informations sur les personnes physiques à partir de données anonymes ou anonymisées, de données publiques ou de données concernant des tierces parties, un droit centré sur la restriction de l'accès aux données personnelles serait condamné à l'échec : l'important serait donc de recentrer le droit sur l'usage qui est fait des données, pour déterminer si celui-ci porte atteinte à la vie privée.

La conception de la vie privée défendue par Wachter & Mittelstadt est une conception holistique, qui comprend notamment un droit de comprendre la manière dont on est perçu (*self-presentation*). Par exemple, l'obligation pour les IAs de s'identifier en tant que telles est évoquée dans les *Lignes directrices*[64] au nom de la transparence, mais on pourrait aussi la voir comme une question de vie privée au titre de cette conception large : une IA implique très souvent une collection de données et un apprentissage, chose que l'utilisateur a le droit de savoir pour protéger sa vie privée comme il a le droit d'être informé de la présence de cookies. Et cela implique aussi une information sur la manière dont nous sommes perçus. Cette question de la compréhension de la manière dont nous sommes perçus, qu'on choisisse ou non de l'inclure dans le respect de la vie privée, est voué à prendre de plus en plus d'importance au fur et à mesure que se multiplient les décisions institutionnelles basées sur des modèles de ML, en finance, assurance, ressources humaines, etc.

Une telle perspective jette une lumière critique sur certaines des approches que nous avons discutées, qu'il s'agisse de la suppression des données personnelles ou des mesures contre les inversions de modèles, dans la mesure où elles sont toujours centrées sur la restriction de l'accès aux données. Si cette perspective ne prive bien sûr pas ces mesures de toute pertinence, elles ne seraient plus placées au centre du paradigme juridique dans les approches défendues par Purtova ou Wachter & Mittelstadt, dont le but n'est plus tant de protéger une catégorie de données que de réglementer l'usage des connaissances inférées sur les personnes.

L'introduction d'un tel « droit à la l'inférence raisonnable » aurait des conséquences juridiques profondes, non seulement en termes de refontes des catégories juridiques, mais en formulations de droit à une inférence scientifiquement valide, en droit à l'explication – qui comprendrait un droit de comprendre la façon dont nous sommes perçus par les institutions- et bien d'autres choses encore. Que changerait-elle pour une approche de respect de la vie privée dès la conception applicable aux dernières évolutions du ML ? Elle diminuerait assurément la pertinence de la classification des modèles de ML invertibles en donnée ou en logiciel : ce qui compte dans cette approche, ce sont les inférences réalisées en bout de course sur les personnes physiques. L'approche par désentraînement des modèles permettrait une évolution technique en accord avec cette philosophie, et qui pourrait contourner les problèmes posés par les autres approches. Elle offre en effet une mise à jour du pouvoir d'inférence des modèles qui permettrait de prendre en compte le droit à l'oubli et autres objections faites à l'égard de certaines inférences du modèle. Cependant, il est encore très tôt pour se prononcer sur sa viabilité technique.

Par sa conception holistique de la vie privée, cette approche interroge également la portée d'une approche respectueuse de cette valeur dès la conception. Pour Wachter & Mittelstadt, le respect de la vie privée va bien plus loin que le respect de l'intimité ou de certaines informations

[64] *Ibid*, p.22.

sensibles. Elle implique un droit de regard sur la façon dont nous sommes perçus, notamment par des institutions utilisant de plus en plus le profilage basé sur le ML pour construire une représentation des individus. Wachter & Mittelstadt insistent à juste titre sur le risque que représente l'opacité grandissante de la perception des individus par les organisations, et de voir le destin des individus déterminés par une identité bureaucratique généré par le profilage à laquelle ils ne pourront pas échapper pour prendre un nouveau départ dans la vie. Sans contester la pertinence de ces questions, on peut se demander s'il est pertinent de les inclure dans la conception de la vie privée, ou s'il vaudrait mieux les inclure dans un droit plus général, comme le « droit à l'autodétermination informationnelle » formulé par la Cour Constitutionnelle allemande. La question doit être reproduite pour l'approche de respect de la vie privée dès la conception : doit-elle affronter des problèmes tels que la capacité à comprendre comment nous sommes perçus par les algorithmes ou le droit à l'oubli ? Ou s'agit-il là d'une extension du concept de vie privée qui risque de lui faire perdre son apport propre, et de créer des confusions avec les autres branches de l'éthique dès la conception comme l'équité ou l'explicabilité dès la conception ?

Conclusion

Ce travail vise à promouvoir le développement de modèles de ML respectueux de la vie privée dans une perspective compatible avec la philosophie libriste, l'esprit des communs numériques et la reproductibilité de la recherche. Dans un premier temps, nous nous sommes inscrits dans une approche respectueuse de la vie privée dès la conception. Notre compréhension de cette approche se démarque d'une volonté de résoudre tous les problèmes éthiques en amont du déploiement, qui serait à la fois technocratique et irréaliste. Nous défendons par contraste une approche éthique qui débute à la conception, mais s'étend sur tout le cycle de vie de l'artefact technique, et inclut d'emblée la possibilité de retour sur expérience formulée par les parties prenantes, essentiel à la fois pour éviter une délégation de tous les problèmes moraux aux développeurs et une confiance excessive dans une approche *ex ante*. Si nous défendons cette vision pour le respect de la vie privée, elle est aussi valide pour toutes les approches éthiques dès la conception, et leur permet d'être à la fois politiquement plus ouvertes et scientifiquement plus robustes.

Dans un deuxième temps, nous avons exposé les dilemmes singuliers auxquels fait face la sécurisation des données personnelles encodées dans les modèles de ML. Nous avons formulé dix principes, qui sont autant de questions posées à la communauté des développeurs en ML. Nous souhaitons avant toute chose lancer une discussion sur le domaine d'applicabilité de tels principes, et sur les raisons précises de cette applicabilité, ou de son absence. Vu l'ampleur de la pratique industrielle du ML, une telle discussion ne peut être menée qu'à partir d'un ample engagement de la communauté, et d'un retour sur expérience des tentatives de développement éthiques dès la conception.

Nos principes sont autant d'options de développement éthiques dès la conception, qui doivent être considérées par les développeurs dans leur ordre de présentation. Nous avons commencé par souligner la nécessité d'une documentation rigoureuse de la finalité du traitement et de son évolution dans le cadre de recherche scientifique, car c'est la finalité du traitement qui détermine l'ampleur de la collecte (principe 1), et guide les réflexions sur les risques d'un pouvoir prédictif trop fin (principe 2) ou de la suroptimisation (principe 3). Nous avons ensuite exploré trois stratégies de développement de modèles de ML : la restriction de la collecte, la lutte contre la rétro-ingénierie, et la vérification publique des modèles.

Malgré les immenses difficultés pratiques qu'elle peut présenter, l'approche par restriction de la collecte se doit d'être considérée en première option, car elle seule permet, dans l'état actuel de l'art, de promettre une sécurisation parfaite de la vie privée en supprimant les données personnelles de la base d'entraînement (principe 4). Outre sa rupture radicale avec les pratiques actuelles de collecte de données, une telle voie devrait aussi faire face aux problèmes de

l'opérationnalisation de la notion de donnée personnelle, problème classique des approches éthiques dès la conception. Ce problème d'opérationnalisation des définitions juridiques peut aussi affecter des stratégies de restriction de la collecte en apparence moins ambitieuses, comme l'approche restreinte aux données sensibles (principe 5) ou l'approche centrée sur le geste de publication (principe 6). L'approche centrée sur le geste de publication pose aussi des problèmes importants de mise à jour des données.

D'un autre côté, une approche fondée sur la sécurisation contre la rétro-ingénierie des modèles aurait pour avantage d'éviter tous les problèmes liés à l'opérationnalisation des notions éthiques, puisqu'elle vise à restreindre l'accès à toutes les données encodées dans le modèle. Cependant, aucune des approches existantes ou naissantes dans la littérature ne permet une sécurisation complète des données. Nous invitons donc les développeurs à considérer cette stratégie comme une deuxième option, et à déclarer les mises prises contre la rétro-ingénierie, et leur position par rapport à l'état de l'art (principe 7), tout en soutenant la diffusion des outils logiciels sous licence libre (principe 8).

Dans ce contexte de sécurisation imparfaite, la vérification publique des propriétés de sécurité, si elle présente de grands avantages en termes de construction de la confiance et de progrès scientifique, pose aussi un dilemme à la culture du logiciel libre, des communs numériques et de la science ouverte, dans la mesure où ces cultures sont toutes attachées à la publicité du logiciel : soit accepter de possibles failles de sécurité considérables en ouvrant les modèles à la vérification publique, soit accepter la délégation du contrôle à un organisme centralisé et restreignant l'accès à l'information, en rupture avec les idéaux de publicité et d'horizontalité très présents dans cette culture. Nous invitons donc les développeurs à prendre leur responsabilité au cas par cas, et à offrir les modèles à la vérification publique des modèles à chaque fois que les contraintes de sécurité le permettent, en détaillant les raisons de leur choix, les mesures prises et leur position par rapport à l'état de l'art (principe 9).

Dans l'état actuel de l'art, il semble qu'il n'existe pas d'approche combinant sécurisation complète, possibilité d'une définition rigoureuse des objectifs et respect des idéaux politiques de publicité et d'horizontalité. Une telle situation, si elle est vouée à perdurer, risque de faire des enjeux de respect des données personnelles en ML un point de contentieux politique dans la communauté des développeurs libristes, des partisans des communs et de la reproductibilité de la recherche. Nous ne pouvons pas trancher ici les profonds problèmes politiques qui vont avec l'abandon des idéaux libristes de transparence et d'horizontalité, ou avec leur défense jusqu'au-boutiste. Nous proposons donc pour finir une exception à notre approche générale, en invitant les développeurs à réfléchir autant aux propriétés techniques du ML qu'aux difficultés institutionnelles qui fondent ces prises d'exception (principe 10).

Pour résumer, notre approche consiste en une première étape visant à un arbitrage raisonnable pour la vie privée entre évolution scientifique des finalités du traitement, extension de la collecte et maîtrise de la puissance prédictive du modèle (principes 1 à 3). Il offre ensuite deux principales options stratégiques : une approche forte par exclusion de certaines données de la base d'entraînement, modulée en trois options (principes 4 à 6), et une approche plus souple basé sur la défense contre la rétro-ingénierie des modèles (principe 7). Remarquons au passage que ces deux options ne sont pas totalement exclusives : on pourrait parfaitement combiner une exclusion de certain type des données avec des mesures contre la rétro-ingénierie des modèles, le manque de maturité des techniques de détection d'informations problématiques pouvant justifier le recours à deux couches de sécurité. Ces deux options peuvent aussi être combinées à des mesures de publicité des outils logiciels et de vérification publique des modèles (principes 8 & 9). Enfin, nous ne pouvons prendre ici position sur le délicat problème des exceptions au respect de la vie privée justifiées par une situation d'urgence, qui peuvent aujourd'hui inclure des modèles fondés sur une collecte en masse des données personnelles, disposant d'un pouvoir prédictif invasif et susceptibles à des attaques de rétro-ingénierie. Nous invitons les développeurs participant à de telles mesures

d'exception à prendre clairement leurs responsabilités, et à offrir une clarification de leur position qui comprenne à la fois une discussion des propriétés propres au ML et une prise en compte des dangers de dérive institutionnelle, notamment par une restriction stricte sur l'accès et l'usage du modèle (principe 10)[65].

Il convient de souligner que ce travail dépend profondément de l'état de l'art en sécurité du ML : s'il existait une solution technique prévenant toute inférence de donnée personnelle à partir d'un modèle de ML, il serait pratiquement sans objet. Les progrès de la confidentialité différentielle en général, et de sa forme fédérée et appliquée dès la conception en particulier, pourraient grandement soulager les différentes inquiétudes soulevées dans cette charte. Le dialogue que nous cherchons à lancer est donc largement un dialogue autour de la frontière mouvante entre les problèmes réglés par les techniques de sécurité et ceux qui ne le sont pas.

Plusieurs relecteurs précoces de ce travail ont suggéré d'intégrer un barème ou une forme de notation montrant le degré de suivi de la charte. Une telle idée tombe a priori sous le sens, puisque les différents principes constituent bien une approche à suivre dans son ordre naturel et non juste une collection de principes sans rapport entre eux. Mais un examen plus rapproché montre qu'elle serait en réalité très dure à implémenter. On pourrait ainsi penser que l'approche par exclusion totale des données personnelles serait la plus forte, et devrait recevoir la note maximale. Mais entre une exclusion des données personnelles imparfaite -comme le seront presque fatalement les outils de la littérature naissante- et une approche qui combinerait par exemple exclusion partielle de données sensibles et défense puissante contre la rétro-ingénierie, il n'est pas du tout évident de déterminer quelle est la plus forte. Par conséquent, si la charte dessine bien un parcours de questions à suivre pour le développeur, elle n'offre pas quelque chose comme une progression homogène qu'on pourrait résumer par une note.

Dans un troisième et dernier temps, nous avons également vu que certains des débats en cours dans la communauté des juristes risquaient d'avoir une grande influence sur le futur des approches respectueuses de la vie privée dès la conception. L'encodage des données personnelles dans les modèles s'inscrit au cœur des problématiques menant à la remise en cause d'un paradigme juridique centré sur la protection des données personnelles. Selon la critique de Purtova, il existe un risque qu'une interprétation maximaliste du droit des données personnelles le mène à devenir un « droit de tout et n'importe quoi » complètement inapplicable. L'évolution future de la notion de donnée personnelle, et son maintien même dans le cadre juridique, dépend donc crucialement du sort à faire à cette interprétation maximaliste, et sa compréhension large du sens auquel une donnée personnelle concerne une personne. Les approches éthiques dès la conception dépendent fortement de ce débat d'interprétation juridique, mais elles pourraient aussi contribuer à le nourrir par leur exploration des limites de l'opérationnalisation des concepts, qui sont aussi des défis d'implémentation du droit dans la pratique informatique.

Selon une autre perspective critique, celle de Wachter & Mittelstadt, le droit actuel n'offre qu'une protection imparfaite dans un monde où la distinction nette entre données et programmes ne peut plus être placée au fondement du droit, qui devrait être recentré sur la notion d'inférence raisonnable. Une telle approche aurait pour mérite de dissoudre les problèmes de catégorisation posés par le ML, dont les modèles peuvent être considérés à la fois comme des données et comme des logiciels au titre du droit actuel. Mais cette approche pose aussi problème par l'ambition de sa conception de la vie privée, qui en vient à inclure le droit à comprendre la perception des individus par les institutions et le droit à l'oubli. Ces questions sur la portée du concept de vie privée se reproduisent pour l'approche dès la conception, et interrogent sa relation aux autres approches éthiques dès la conception.

---

[65]En accord avec notre philosophie de travail consistant à aller aux limites de l'état du droit, nous ne mentionnons pas une possible obligation de détruire les données une fois leur rôle d'urgence accompli. En plus de limites temporelles dures pour certains usages, le RGPD contient une exigence de définition d'un temps de conservation nécessaire des données en fonction de la finalité de l'usage (voir Considérant 39 & Article 5-e).

Remerciements

Ce travail est parti du désir de membres du laboratoire d'informatique de l'Université de Lorraine, le LORIA, de discuter des enjeux éthiques des modèles d'IA dans le cadre du projet OLKi conduit avec les Archives Henri Poincaré. Il est particulièrement redevable à la volonté de Christophe Cerisera, directeur du laboratoire de Traitement Automatique de la Langue, de discuter les enjeux de vie privée soulevés par la collecte des données linguistiques et la production des modèles pour cette discipline. La délicate construction de l'objet de cette charte et de sa structure provient avant tout des nombreuses conversations que nous avons eues ensemble.

Judith Rochfeld et Aurélien Bellet ont eu la générosité de relire ce document pour revoir les parties touchant à leur spécialité, soit respectivement l'état de l'art juridique en droit des données et la sécurité en ML. Il va de soi que leur aide a été immensément bénéfique. Les commentaires de Bastien Guerry sur une des premières versions de ce document ont grandement aidé à en faire ressortir les hypothèses les plus fondamentales.

Je tiens enfin à remercier mes supérieurs hiérarchiques Cyrille Imbert et Anna Zielinska pour leur aide dans la conception de cette charte, leur travail de relecture et pour la confiance qu'ils m'ont accordée. L'aide administrative toujours prévenante d'Aurore Coince a joué un rôle aussi discret qu'essentiel.

Ce travail a bénéficié d'une aide de l'État, gérée par l'Agence Nationale de la Recherche, au titre du projet Investissements d'Avenir Lorraine Université d'Excellence, portant la référence ANR-15-IDEX-04-LUE. Les dernières retouches ont eu lieu au début de ma prise de fonction postdoctorale dans le *Carl Friedrich von Weiszäcker Zentrum* de l'Université de Tübingen.