UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ADMISSIBILITÉ DE PREUVES ISSUES DE TECHNIQUES D'APPRENTISSAGE AUTOMATIQUE EN DROIT CRIMINEL CANADIEN

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

MAÎTRISE EN DROIT

PAR

MARIANNE OZKAN

SEPTEMBRE 2022

REMERCIEMENTS

Je tiens à remercier M. Hugo Cyr pour les conversations stimulantes sur le droit, bien sûr, mais aussi sur la politique et les nouvelles technologies en général, de l'intelligence artificielle aux chaînes de blocs. M. Cyr incarne pour moi l'idéal académique. C'est l'image que je retiens de mes années passées à la faculté de droit, de même que celles des professeurs qui m'ont marquée, Olivier Barsalou et Alejandro Lorite Escorihuela. Je remercie ma famille qui se reconnaîtra dans ce qui suit : les discussions philosophiques dès l'aube, l'engouement devant les sujets de recherche possibles, les questions statistiques à élucider. Vous m'avez à la fois rassuré et secoué.

DÉDICACE

À mes parents dont la lumière guide mes pas.

TABLE DES MATIÈRES

LISTE DES FIGURES viii
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES x
RÉSUMÉ xi
ABSTRACT xii
INTRODUCTION 1
CHAPITRE I L'ADMISSIBILITÉ DE LA PREUVE SCIENTIFIQUE EN DROIT 10
1.1 L'admission de la preuve scientifique en tant que témoignage d'expert 10
1.2 La démarche d'évaluation de l'admissibilité de la preuve scientifique 11
1.2.1 La première étape: les quatre critères de l'arrêt *Mohan* 12
1.2.2 Deuxième étape : l'analyse des avantages versus les risques en cas d'inclusion de la preuve 15
1.2.3 La fiabilité de la preuve scientifique 17
1.3 Résumé de la démarche d'admissibilité de la preuve scientifique 31
CHAPITRE II L'APPRENTISSAGE AUTOMATIQUE 33
2.1 Qu'est-ce que l'apprentissage automatique 33
2.1.1 Les modes d'apprentissage 38
2.1.2 Les composantes d'un système d'apprentissage automatique 41
2.1.3 Les étapes de production d'un système d'apprentissage automatique 43
2.1.4 Exemples d'apprentissage automatique 44
2.1.5 L'intelligence artificielle au Canada 46
2.2 Résumé 49
CHAPITRE III LA FIABILITÉ DU PRINCIPE FONDAMENTAL 50
3.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de son principe fondamental 50
3.2 L'approche statistique de l'apprentissage automatique 50
3.2.1 Déduction versus induction 50
3.2.2 L'approche statistique classique versus l'apprentissage automatique 52
3.2.3 Les enjeux liés à l'utilisation des statistiques 55

3.3 Le modèle du système 65
3.3.1 Les enjeux du modèle 65
3.4 Résumé 67
CHAPITRE IV LA FIABILITÉ DU PROCESSUS 68
4.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de son processus 68
4.2 Les enjeux 68
4.2.1 Les données 68
4.2.2 L'intervention humaine 71
4.3 Résumé 73
CHAPITRE V L'EXPLICATION DU RÉSULTAT 74
5.1 Les questions que les juristes doivent se poser devant un système en lien avec l'explication du résultat obtenu 74
5.2 L'explication algorithmique 75
5.3 Les fonctions de l'explication : de la fiabilité du système au processus judiciaire 76
5.4 Les enjeux de l'explication 78
5.4.1 Complexité des modèles 78
5.4.2 Les fausses corrélations 79
5.4.3 Les coûts et le secret commercial 81
5.5 L'intelligence artificielle explicable 83
5.5.1 Portée de l'explication 83
5.5.2 Caractéristiques souhaitables de l'explication 84
5.5.3 Exemples d'explication 86
5.6 Résumé 92
CHAPITRE VI LA FIABILITÉ DE L'EXPLICATION 93
6.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de l'explication 93
6.2 Expliquer et interpréter un modèle : une affaire de précision 94
6.3 Les modèles complexes 95
6.4 Critères d'évaluation de l'explication 103
6.5 Résumé 105
CHAPITRE VII LA PERFORMANCE DU SYSTÈME 106

7.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la performance du système 106

7.2 La performance globale 107

7.2.1 Les métriques 107

7.3 Les enjeux de performance 114

7.3.1 La performance par sous-groupes 114
7.3.2 La discrimination statistique 114
7.3.3 Métriques de détection : l'effet disproportionné et l'effet disproportionné conditionnel 123
7.3.4 Le choix des métriques selon le contexte 126

7.4 Résumé 129

CHAPITRE VIII ASSURANCE QUALITÉ 130

8.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la scientificité des tests employés 130

8.2 Les enjeux reliés aux tests 130

8.2.1 La représentativité des échantillons 131
8.2.2 La reproductibilité 132
8.2.3 La répétabilité (*replicability*). 134
8.2.4 Autonomie et opacité des modèles 135

8.3 Assurance qualité 136

8.3.1 Standards et normes 138
8.3.2 Revue et audits 138

8.4 Résumé 142

CHAPITRE IX LA RÉGLEMENTATION 143

9.1 Support réglementaire à l'évaluation de la fiabilité 143

9.2 Recommandations de la communauté juridique 144

9.3 Europe 147

9.3.1 Le Règlement général sur la protection des données (RGPD) 147
9.3.2 La Loi pour une république numérique en France 150
9.3.3 Proposition de règlement du Parlement européen et Résolution du Parlement européen sur l'utilisation de l'intelligence artificielle en droit pénal 152

9.4 États-Unis 156

9.4.1 The Algorithmic Accountability Act 156

9.5 Au Canada 159

9.5.1 Lois sur la preuve au Canada et au Québec 159
9.5.2 La Directive sur la prise de décision automatisée au Canada..... 161
9.5.3 La Loi sur la protection des renseignements personnels 162
9.5.4 La Loi modernisant des dispositions législatives en matière de protection des renseignements personnels du Québec..... 165

CHAPITRE X EXEMPLE D'ÉVALUATION DE LA FIABILITÉ D'UN OUTIL DE RECONNAISSANCE DU LOCUTEUR..... 167

10.1 La reconnaissance du locuteur 167

10.2 Les témoignages auditifs..... 170

10.2.1 La fiabilité du témoignage profane..... 171
10.2.2 Les outils du témoignage expert..... 172

10.3 Les enjeux de la reconnaissance par la voix..... 175

10.4 Le cas..... 177

10.4.1 Fiabilité du principe fondamental..... 178
10.4.2 Fiabilité de la méthode..... 179
10.4.3 Fiabilité de l'explication..... 181
10.4.4 Questions sur la performance 181
10.4.5 Questions sur les tests..... 183
10.4.6 Conclusions sur l'admissibilité de la preuve 183

CHAPITRE XI RÉSUMÉ DES CONDITIONS FAVORABLES À LA FIABILITÉ 187

CONCLUSION..... 188

BIBLIOGRAPHIE..... 194

LISTE DES FIGURES

Figure Page

Figure 1 - Programmation traditionnelle et apprentissage automatique 37

Figure 2 - Exemples d'algorithmes d'apprentissage automatique selon trois grandes approches d'apprentissage : supervisé, non-supervisé et par renforcement 39

Figure 3 - Équivalence verbale du rapport de vraisemblance 57

Figure 4 - Image de l'outil « What-If Tool » de Google 90

Figure 5 - Relation linéaire positive 97

Figure 6 - Relation linéaire négative 97

Figure 7 - Relation non-linéaire 97

Figure 8 - Relation monotone 97

Figure 9 - Relation continue 97

Figure 10 - Relation discontinue 97

Figure 11 - Arbre de décision d'un système de prêt bancaire 101

Figure 12 - Matrice de confusion..... 109

Figure 13 - Deux courbes normales..... 110

Figure 14 - Seuil de décision 110

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

LPRPDE : Loi sur la protection des renseignements personnels et les documents électroniques

RGPD : Règlement général sur la protection des données

RÉSUMÉ

Ce mémoire se penche sur l'un des effets de l'émergence d'outils d'intelligence artificielle sur la pratique du droit. En particulier, nous traitons de l'admissibilité de la preuve issue d'outils utilisant la technique de l'apprentissage automatique, une branche de l'intelligence artificielle. Nous cherchons à établir la fiabilité de cette technique pour fins d'admissibilité en tant que preuve. Nous débutons en cernant la notion de fiabilité d'une preuve scientifique en droit canadien. Nous abordons ensuite les composantes et le fonctionnement de l'apprentissage automatique. Nous analysons les divers aspects de sa fiabilité en soulevant ses vulnérabilités, ce qui nous permet de dégager les conditions propices à la fiabilité de la technique. Nous recensons les instruments légaux qui imposent ou renforcent ces conditions et terminons avec une illustration concrète d'un témoignage expert sur une telle preuve, soit le cas d'un outil visant à cerner l'identité d'un locuteur. Notre démarche nous incite à remettre en question le rôle du tribunal dans l'établissement de la fiabilité d'un outil d'apprentissage automatique, une tâche qui défavorise l'inculpé.

Mots clés : Intelligence artificielle, apprentissage automatique, admissibilité, preuve scientifique, fiabilité, reconnaissance du locuteur.

ABSTRACT

This document concerns one of the impacts of artificial intelligence on the practice of law. In particular, it focuses on the admissibility of results from machine learning techniques as evidence in criminal courts, machine learning being a subset of artificial intelligence. Our goal is to determine the reliability of machine learning techniques in the context of their admissibility in court. To do so, we start by laying out the Canadian courts' method for determining the reliably of technical evidence. We then proceed to describe the components and inner workings of machine learning. Next, we analyse the multiples aspects of machine learning reliability, shedding light on its vulnerabilities and thus on the conditions which favor reliability. We then survey the regulatory instruments which promote or enforce these conditions. We sum-up our analysis with a concrete case involving an expert witness testifying on the reliability of a machine learning speaker identification tool. Our findings question the role of the court in evaluating reliability of machine learning tools, as it may lead to prejudice against the suspect.

Keywords : artificial intelligence, machine learning, admissibility, scientific evidence, reliability, speaker identification.

INTRODUCTION

L'intelligence artificielle est une branche de l'informatique dont l'objectif est de reproduire les facultés humaines de l'intelligence, par exemple la capacité à la résolution de problème ou la créativité[1]. L'intelligence artificielle est aujourd'hui utilisée avec succès dans des domaines aussi variés que la médecine, par exemple pour détecter une pneumonie à partir d'une radiographie[2]; le domaine de la finance, pour la gestion de portefeuille d'actions cotées en bourses[3]; le domaine des ressources humaines pour le recrutement de personnel[4]; le domaine de l'éducation comme aide au

---

[1] BJ Copeland, « Artificial Intelligence » dans *Encyclopedia Britannica*, 2021.

[2] Erik Ranschaert, « Artificial Intelligence in Radiology: Hype or Hope? » (2018) 102:S1 Journal of the Belgian Society of Radiology 20.

[3] Yang Lu, « Artificial Intelligence: A Survey on Evolution, Models, Applications and Future Trends » (2019) 6:1 Journal of Management Analytics 1; Bastien Collard, *L'impact de l'intelligence artificielle dans la gestion de portefeuilles* (mémoire de M. Sc, Université de Louvain, 2019) [non publié].

[4] Nishad Nawaz, « How Far Have We Come With The Study Of Artificial Intelligence For Recruitment Process » (2019) 8:7 International Journal of Scientific & Technology Research 488.

processus d'admission d'étudiants[5] ; le domaine du transport avec les voitures autonomes[6]; ou encore le domaine de la domotique avec les assistants vocaux[7].

Le droit n'est pas en reste[8]. Les récents développements des techniques juridiques ou « legaltech » ont entraîné la création de nombreux outils d'aide à la résolution de conflits légaux[9]. Ils permettent d'analyser rapidement des milliers de documents de jurisprudence afin d'y relever automatiquement les arguments porteurs dans une cause[10] ou en vue de déterminer les probabilités de gagner ou perdre une cause[11]. Le grand public est également ciblé : visant à accroître l'accessibilité à la justice, la

---

[5] Parcoursup, « Parcoursup: Entrez dans l'enseignement supérieur », (2021), en ligne: *Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation* <https://parcoursup.fr/>; Sarah Sermondadaz, « L'algorithme de Parcoursup décrypté par les deux chercheurs qui l'ont conçu », (12 juin 2018), en ligne: *Sciences et Avenir* <https://www.sciencesetavenir.fr/high-tech/informatique/bac-2018-l-algorithme-de-parcoursup-explique-par-les-deux-chercheurs-qui-l-ont-concu_124407>.

[6] Christiana Varnava, « Technology Takes the Wheel » (2019) 2:8 Nature Electronics (Nature Publishing Group) 319.

[7] Lu, « Artificial intelligence », *supra* note 3; Matt Muldoon, « L'intelligence artificielle vocale : qu'est-ce que c'est ? », (24 mars 2021), en ligne: *ReadSpeaker AI* <https://www.readspeaker.ai/fr/blog/what-is-multilingual-neural-text-to-speech-and-how-can-it-help-your-business-12-copy-copy/>.

[8] Mordor Intelligence, « AI Software Market in Legal Industry - Growth, Trends, Covid-19 Impact, and Forecasts (2021 - 2026) », (2020), en ligne: <https://www.mordorintelligence.com/industry-reports/ai-software-market-in-legal-industry>.

[9] Julius Remmers, « Legal Service by Automated Legal Software and Its Legal Impacts, in Particular under the German Act On Out-of-Court Legal Services » (2018) SSRN, en ligne: <https://papers.ssrn.com/abstract=3491347> à la p 5. Voir la figure 1.1.

[10] Casetext, « Casetext Research: Best Legal Research Software », (2021), en ligne: *Casetext* <https://casetext.com/research/>.

[11] Blue J, « Predictive Tax and Employment Law Software », (2021), en ligne: <https://www.bluej.com/ca>.

recherche se penche sur la prise automatisée de décisions juridiques, en remplacement des juges[12]. Ces outils mis à disposition du public ont pour objectif de réduire les coûts et délais associés aux procédures légales[13].

Les acteurs du droit voient également s'infiltrer des outils d'intelligence artificielle au sein de tribunaux, soit parce qu'ils sont contestés[14], par exemple pour leur atteinte à la vie privée[15], soit parce que leurs résultats servent de preuve lors d'un recours en justice[16]. En effet, les résultats d'outils d'intelligence artificielle peuvent servir à étayer les arguments d'une partie en litige. On pense à l'utilisation d'outils prédictifs[17] qui évaluent la probabilité d'un évènement futur, par exemple l'évaluation de risques de

---

[12] Canada, Ministère de la Justice, « Automatisation de la justice - Tendances en matière de justice 2 : Automatisation de la justice. Un aperçu de l'avenir des technologies dans le système judiciaire », (2 avril 2008), en ligne: *Gouvernement du Canada* <https://www.justice.gc.ca/fra/pr-rp/jr/tmj2-jt2/p3.html> à la section « Tribunaux automatisés »; Matt Gooding, « "Robo-Lawyer" Can Predict How Likely You Are to Win a Case », (20 juin 2017), en ligne: *CambridgeshireLive* <www.cambridge-news.co.uk/business/technology/cambridges-lawbot-robo-lawyer-can-13209877>.

[13] Remmers, *supra* note 9 à la p 8.

[14] Robin Allen et Dee Masters, « French Parcoursup Decision », (16 avril 2020), en ligne: *AI Lawhub* <https://ai-lawhub.com/2020/04/16/french-parcoursup-decision/>. À la section « Latest News », quelques cas sont mentionnés sous le titre « Case Law ». L'affaire concernant le logiciel Parcoursup en France est l'un d'eux.

[15] *R (Bridges) v CCSWP and SSHD*, [2019] EWHC 2341 (Admin) . Cette affaire concerne un logiciel de reconnaissance de visage utilisé au Royaume-Uni.

[16] Nicole Hong, « Court to Rule on Voice Analysis in Terrorism Trial », *Wall Street Journal* (11 mai 2015), en ligne: <https://www.wsj.com/articles/court-to-rule-on-voice-analysis-in-terrorism-trial-1431361065>. Ce cas aux États-Unis concerne un logiciel de reconnaissance du locuteur proposé en preuve.

[17] Matthias Leese, Mareile Kaufmann et Simon Egbert, « Predictive Policing and the Politics of Patterns » (2019) 59 British Journal of Criminology 674. L'article concerne les outils de police prédictives qui orientent le déploiement des forces de l'ordre.

récidive d'un détenu qui pourra aider à justifier une sentence[18]. Un autre cas d'utilisation concerne les outils permettant de reconstituer le fil des évènements passés[19]. Par exemple, les tests d'ADN ou l'alcotest, sont aujourd'hui des techniques bien connues des services policiers et sont régulièrement soumis en preuve au tribunal pour établir des faits, par exemple pour établir qu'une personne se trouvait à un endroit précis selon l'analyse génétique de traces trouvées à cet endroit[20] ou, par exemple, pour établir le taux d'alcool d'un conducteur automobile lors de son arrestation[21]. Parmi les nouveaux outils d'intelligence artificielle disponibles aux services policiers, notons la reconnaissance faciale qui permet d'identifier une personne en recoupant une image prise à l'aide d'une caméra de sécurité avec une banque de données de photos

---

[18] Equivant, « Northpointe Suite Risk Need Assessments », (mai 2021), en ligne: *Equivant* <www.equivant.com/northpointe-risk-need-assessments/>.

[19] Rafael Encinas De Munagorri, « Les problèmes de preuve posés par l'évolution des sciences et des technologies » dans *Applied Ethics at the Turn of Millenium*, 2001 : « Les sciences et les technologies peuvent éclairer les faits controversés et faciliter le rôle du juge lorsqu'il doit prendre une décision ».

[20] Marie Angèle Grimaud, « Les enjeux de la recevabilité de la preuve d'identification par ADN dans le système pénal canadien » (1994) 24:2 RDUS 293.

[21] Canada, Commission de réforme du droit, *Les méthodes d'investigation scientifique*, document de travail 34, 1984 [Commission de réforme du droit, « Méthodes d'investigation »] à la section « Analyse d'haleine ».

identifiées[22], ou les outils qui permettent d'identifier une personne à partir d'un mélange complexe d'ADN[23], ou encore à partir d'un enregistrement de sa voix.

C'est cette dernière utilisation de l'intelligence artificielle qui nous intéresse dans ce mémoire, soit l'utilisation du résultat d'un outil d'intelligence artificielle pour établir un élément de preuve lors d'un procès. En particulier, nous nous intéressons aux techniques d'un sous-ensemble de l'intelligence artificielle appelé *apprentissage automatique*. L'apprentissage automatique est une spécialisation de l'intelligence artificielle qui se démarque de l'informatique classique par sa méthode de résolution de problème basée sur l'apprentissage : la machine « apprend » à accomplir une tâche plutôt que de suivre une série d'instructions définies par l'humain[24]. Cette caractéristique en fait une technique unique puisque ce n'est plus l'humain qui dicte le fin détail des instructions mais la machine qui les découvre d'elle-même. C'est une technique relativement nouvelle[25] qui a, de ce fait, peu été abordée dans le contexte du droit de la preuve.

---

[22] Kashmir Hill, « The Secretive Company That Might End Privacy as We Know It », *The New York Times* (18 janvier 2020), en ligne: <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>. L'un des logiciels utilisés par les forces de l'ordre est Clearview AI.

[23] Andrea Roth, « Machine Testimony » (2017) 126:7 Yale LJ, en ligne: <https://digitalcommons.law.yale.edu/ylj/vol126/iss7/1>. L'auteure se penche sur True Allele, un outil de détection d'ADN.

[24] Selmer Bringsjord et Naveen Sundar Govindarajulu, « Artificial Intelligence » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, summer 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

[25] Cédric Villani et al, *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*, 2018 aux pp 9-10.

Le droit de la preuve établit les normes d'admissibilité d'un élément en tant que preuve au tribunal. Parmi celles-ci, une preuve issue d'une nouvelle technique se doit d'être évaluée en fonction de sa *fiabilité*[26]. Il n'y a pas de « grille d'analyse » unique permettant d'évaluer la fiabilité d'une preuve scientifique en droit canadien[27]. Selon le contexte, l'examen retiendra divers critères[28]. La fiabilité exige néanmoins que la technique réponde aux exigences de la méthode scientifique[29], laquelle pourra être évaluée selon les critères identifiés dans l'arrêt *J.-L.J.* [30] qui constitue une référence en droit canadien[31]. Ces critères sont les suivants : la technique doit avoir été testée, son taux d'erreurs doit être connu, elle doit avoir été revue par les pairs (par exemple, par le biais de publications scientifiques) et faire l'objet d'une acceptation générale auprès

---

[26] *R c Mohan*, [1994] 2 RCS 9, p. 25; 1994 CanLII 80 [*Mohan*] : « la preuve d'expert qui avance une nouvelle théorie ou technique scientifique est soigneusement examinée pour déterminer si elle satisfait à la norme de fiabilité et si elle est essentielle en ce sens que le juge des faits sera incapable de tirer une conclusion satisfaisante sans l'aide de l'expert. Plus la preuve se rapproche de l'opinion sur une question fondamentale, plus l'application de ce principe est stricte ».

[27] Patenaude, Pierre, « Modern Scientific Evidence » (2000) 30:2 RDUS 407-417 à la p 3 : « The Supreme Court has not established any grille d'analyse to determine the validity and probative value of novel scientific evidence »; P Brad Limpert, « Beyond the Rule in Mohan: A New Model for Assessing the Reliability of Scientific Evidence » (1996) 54:1 UT Fac L Rev 65. Limpert citant le Juge Sopinka dans The Law of Evidence in Canada à la note 102 : « English and Canadian courts have not attempted to create a single standard of admissibility [for novel scientific evidence] ».

[28] *R v Johnston*, [1992] 1992 CanLII 12790 (ON SC) [*Johnston*]. Par exemple, cet arrêt compte plus de 12 critères.

[29] David Paciocco, « L'évaluation du témoignage d'opinion pour en établir l'admissibilité : les leçons récentes du droit de la preuve » (1995) 26:3 RGD 425. La fiabilité a trait à la « validité scientifique ».

[30] *R c J-LJ*, [2000] 2000 CSC 51, [2000] 2 RCS 600 [*J.-L.J.*].

[31] Gabriel Stettler, « L'administration de la preuve scientifique en droit Nord-Américain. » (2019) 97:1 Can Bar Rev 177.

de la communauté scientifique[32]. En droit criminel canadien, étant donné les enjeux pour l'accusé, le droit « exige une preuve hors de tout doute raisonnable avant de condamner, et […] le moindre doute doit profiter à l'accusé »[33].

Le présent document se penche sur l'admissibilité en preuve des techniques d'apprentissage automatique dans le contexte du droit criminel canadien. Pour ce faire, nous retiendrons les critères énoncés dans *J.-L.J.* Nous évaluerons également deux aspects indissociables de la fiabilité. Le premier aspect est le principe fondamental de la technique, c'est-à-dire les prémisses sur lesquelles elle repose, tel que recommandé par la Cour suprême du Canada dans *R. c. Bingley*[34]. Le deuxième aspect est l'explication du résultat obtenu, car c'est par l'entremise de cette explication que les résultats sont communiqués au juge des faits[35], tâche qui peut s'avérer un défi dans le cas de certains systèmes d'apprentissage automatique[36]. Le tout est mis en œuvre dans un cas d'étude fictif d'une preuve par reconnaissance vocale.

Ce document pourra servir de guide à quiconque désire démystifier l'apprentissage automatique car il explique son fonctionnement et ses diverses composantes. Il met

---

[32] *J.-L.J.*, *supra* note 30 au para 33.

[33] Grimaud, *supra* note 20 à la p 326.

[34] *R c Bingley*, 2017 CSC 12, [2017] 1 RCS 170 au para 43 [*Bingley*]. Pour être admissible en preuve, « les principes scientifiques sous-jacents [à la nouvelle technique doivent être] suffisamment fiables ».

[35] *R v Béland*, [1987] [1987] 2 SCR 398, 1987 CanLII 27 au para 24. C'est en effet par l'entremise de l'explication par le témoin expert que les résultats, « leur nature et leur sens sont communiqués au juge des faits ».

[36] Andrew D Selbst et Solon Barocas, « The Intuitive Appeal of Explainable Machines » (2018) 87 Fordham L Rev 1085 à la p 1094. L'apprentissage automatique peut engendrer des systèmes complexes au point d'être inintelligibles, ce qui pose des défis quant à l'explication des résultats.

également en lumière les enjeux techniques de l'apprentissage automatique, toujours dans l'optique d'évaluer sa fiabilité. Il met donc en relief les vulnérabilités de la technique quant à l'incertitude de ses résultats. Le document pourra également servir de canevas concret pour interroger un témoin expert lors de la présentation de la preuve. En effet, le document propose des questions à adresser à l'expert pour évaluer la fiabilité de la preuve sous ses différentes facettes.

Le mémoire est structuré comme suit. Nous débuterons par réviser la démarche d'admissibilité de la preuve en droit canadien (chapitre 1). Nous nous pencherons ensuite sur la technique de l'apprentissage automatique dans son ensemble (chapitre 2). Ensuite, nous analyserons les divers aspects de la fiabilité de l'apprentissage automatique, en soulevant les enjeux qui s'y rattachent. Nous débuterons par le principe fondamental de la technique (chapitre 3). Ensuite, nous aborderons le processus de production d'un résultat issu de l'apprentissage automatique, notamment en regard de sa matière première, soit les données consommées par un outil d'apprentissage automatique (chapitre 4). Nous nous pencherons sur l'explication du résultat obtenu (chapitre 5) et nous évaluerons la fiabilité de cette explication et son incidence sur la valeur probante de la preuve (chapitre 6). Nous examinerons le critère de l'incertitude du résultat qui permet d'évaluer la fiabilité du système en termes de taux d'erreurs dans le chapitre consacré à la performance (chapitre 7). Finalement, nous examinerons dans le chapitre consacré à l'assurance qualité les critères *J.-L.J* [37] concernant les tests ainsi que la revue par les pairs (chapitre 8).

Au début de chaque chapitre, nous suggérons des questions à adresser au témoin expert. Celles-ci pourront contribuer à évaluer la fiabilité de la preuve en cour. Ces

---

[37] *J.-L.J.*, *supra* note 30.

questions cherchent à mettre en relief le traitement des enjeux techniques spécifiques à la fiabilité de l'apprentissage automatique, enjeux que nous exposerons tout au long du mémoire. Ainsi, selon les réponses de l'expert sur la prise en compte de ces enjeux, nous pourrons mieux jauger la valeur probante de la preuve.

Nous abordons les aspects réglementaires qui touchent la fiabilité des systèmes d'apprentissage automatique (chapitre 9). Ceux-ci répondent aux enjeux techniques soulevés et pourront épauler une démarche d'évaluation de la fiabilité d'une preuve issue d'un système par apprentissage automatique en cour. Nous terminerons avec un cas pratique concernant un système d'identification du locuteur où nous simulerons un témoignage d'expert (chapitre 10). Ce cas nous permettra d'aborder l'ensemble des éléments discutés dans ce mémoire qui s'appliquent. De même, ce cas nous permettra d'aborder le quatrième critère de l'arrêt *J.-L.J* [38], soit l'acceptation générale puisqu'il est plus pertinent de traiter ce critère dans le cadre d'analyse d'un outil précis que de la technique générale de l'apprentissage automatique. Finalement, nous conclurons sur un constat par rapport au rôle des tribunaux dans l'évaluation de la fiabilité d'un système d'apprentissage automatique, rôle qui selon nous défavorise l'accusé en augmentant les risques de minimiser le doute raisonnable et en augmentant indûment le poids procédural qui lui est imposé.

---

[38] *Ibid*.

CHAPITRE I

L'ADMISSIBILITÉ DE LA PREUVE SCIENTIFIQUE EN DROIT

Dans ce chapitre nous abordons la démarche d'admissibilité d'une preuve scientifique en cour et décrivons comment l'aspect de la « fiabilité » de la preuve s'inscrit dans cette démarche.

1.1 L'admission de la preuve scientifique en tant que témoignage d'expert

Au Canada, tant en vertu de la common law que du droit civil québécois, il existe deux types de témoignages admissibles en cour, soit le témoignage ordinaire et le témoignage d'opinion[39]. Dans le premier cas, le témoin ne doit parler que de ce qu'il a directement observé[40]. Dans le deuxième cas, le témoin fournit à la Cour une opinion et donc « peut témoigner sur des faits dont il n'a pas une connaissance personnelle »[41]. Le témoignage d'opinion peut être livré par un témoin ordinaire seulement si l'opinion

---

[39] Nicolas Bellemare, « Chapitre VII - La preuve pénale » dans *Droit pénal: procédure et preuve,* vol 12, Montréal, Éditions Yvon Blais, 2019 à la p 149; *CcQ* art 2843; *Loi sur la preuve au Canada*, LRC 1985, c C-5; Canada, Institut national de la magistrature*, Manuel scientifique à l'intention des juges canadiens,* 2018 [Manuel scientifique] à la p 20.

[40] Bellemare, *supra* note 39 à la p 149.

[41] *Ibid*.

fait partie du sens commun et qu'il « est à peu près impossible de séparer les inférences du témoin des faits sur lesquelles elles se fondent »[42]. Le témoignage d'opinion doit être livré par un expert lorsque la preuve nécessite des connaissances particulières qui ne font pas partie des connaissances usuelles d'un citoyen ordinaire[43]. En effet,

> [l]e droit reconnaît que, dans la mesure où les questions exigent des connaissances ou des compétences particulières, les juges et les jurés ne sont pas forcément en mesure de tirer une véritable conclusion d'après les faits relatés par les témoins. Le témoin est par conséquent admis à faire part de son opinion sur ces questions, pourvu qu'il soit un expert en la matière[44].

La preuve scientifique est donc admissible en tant que témoignage d'opinion par un expert pourvu que certains critères soient respectés[45]. Nous abordons en détail ces critères dans le présent chapitre.

1.2 La démarche d'évaluation de l'admissibilité de la preuve scientifique

L'admissibilité du témoignage d'expert est évaluée en deux étapes selon la Cour suprême dans l'arrêt *Bingley*[46]:

---

[42] *Graat c La Reine*, [1982] 2 RCS 819 à la p 824.

[43] Bellemare, *supra* note 39 aux pp 149, 150.

[44] *White Burgess Langille Inman Ltd c Abbott and Haliburton Co Ltd*, 2015 CSC 23, [2015] 2 RCS 182 au para 15 [*White Burgess*], citant le professeur Tapper.

[45] David M Paciocco, *The Law of Evidence*, 8e éd, Toronto, Ontario, Irwin Law, 2020 à la section 4.2.

[46] *Bingley*, *supra* note 34.

> Premièrement, celui-ci [le témoignage] doit satisfaire aux quatre critères énoncés dans l'arrêt *Mohan* [47]: pertinence, nécessité, absence de toute règle d'exclusion et expertise particulière. Deuxièmement, le juge du procès doit soupeser les risques éventuels et les avantages que présente l'admission du témoignage.

1.2.1 La première étape: les quatre critères de l'arrêt *Mohan*

L'arrêt *Mohan*, rédigé par le juge Sopinka en 1994, a établi un tournant au sein des tribunaux canadiens en ce qui concerne l'admissibilité d'une preuve scientifique[48]. Auparavant, les tribunaux utilisaient des critères variables. L'un d'eux était le critère de « l'acceptation générale », critère inspiré du droit américain où il avait été largement adopté suite à l'affaire *Frye c. United States* en 1924[49]. En vertu de ce critère, le juge devait rejeter toute preuve qui n'avait pas été confirmée par la communauté scientifique. Si ce critère avait l'avantage de prémunir le procès contre les « sciences de pacotilles » [50], il avait également le désavantage de reléguer le juge au deuxième plan, effacé derrière la communauté scientifique[51].

L'arrêt *Mohan*, dans un verdict unanime de la Cour, en 1994, mit un terme à cette pratique. L'arrêt redonna au juge son rôle de premier plan comme « gardien du processus judiciaire » visant à éloigner du procès les témoignages pseudoscientifiques.

---

[47] *Mohan*, *supra* note 26.

[48] Stettler, *supra* note 31 au para 43.

[49] *Frye v United States*, 293 F 1013, 1923 US App Lexis 1712 [*Frye*].

[50] Stettler, *supra* note 31 au para 18 : « Ces pseudosciences définies comme des "théories ou croyances qui sont nouvelles ou spéculatives et qui manquent de bases scientifiques" procèdent du charlatanisme le plus flagrant aux fraudes scientifiques les plus élaborées ».

[51] *Ibid* au para 37.

Les critères d'admissibilité de l'arrêt *Mohan* constituent aujourd'hui le socle de l'évaluation de l'admissibilité de la preuve scientifique en droit canadien[52]. Ils ont été confirmés par la Cour suprême du Canada en 2015 dans l'arrêt *White Burgess Langille Inman c Abbott and Haliburton Co*[53].

Les critères de l'arrêt *Mohan* sont décrits dans les quatre sections suivantes.

1.2.1.1 La pertinence

La preuve est pertinente lorsqu'elle est « liée au fait concerné qu'elle tend à établir »[54]. Dans l'arrêt *R c J.-L.J.*, la Cour suprême précise que la preuve est pertinente « lorsque, selon la logique et l'expérience humaine, elle tend jusqu'à un certain point à rendre la proposition qu'elle appuie plus vraisemblable qu'elle ne le paraîtrait sans elle »[55].

1.2.1.2 La nécessité d'aider les juges des faits

L'expert doit fournir des conclusions aux juges et jurés qu'eux-mêmes ne pourraient tirer à partir des faits, faute de connaissances :

> L'opinion d'un expert est recevable pour donner à la cour des renseignements scientifiques qui, selon toute vraisemblance, dépassent l'expérience et la connaissance d'un juge ou d'un jury. Si, à partir des faits

---

[52] *Ibid* aux para 43-44.

[53] *White Burgess*, *supra* note 45; *Manuel scientifique*, *supra* note 39 à la p 21.

[54] *Mohan*, *supra* note 26 à la section « III - Analyse ».

[55] *J.-L.J.*, *supra* note 30 au para 47.

> établis par la preuve, un juge ou un jury peut à lui seul tirer ses propres conclusions, alors l'opinion de l'expert n'est pas nécessaire[56].

1.2.1.3 L’absence de toute règle d’exclusion

La preuve ne doit pas contrevenir à une règle distincte de la règle de preuve d’opinion. L’arrêt *Mohan* cite en exemple le cas *R. c. Morin* [[57]] où la preuve, malgré le fait qu’elle respectait les critères de la preuve d’opinion, fut jugée inadmissible « sur le fondement de la règle qui interdit au ministère public de produire une preuve de la propension de l'accusé à moins que ce dernier n'ait mis sa moralité en jeu »[58].

1.2.1.4 L’expertise particulière

L’expert doit posséder des qualifications particulières. L’arrêt *Mohan* précise que « la preuve doit être présentée par un témoin dont on démontre qu'il ou elle a acquis des connaissances spéciales ou particulières grâce à des études ou à une expérience relative aux questions visées dans son témoignage »[59].

L’arrêt *White Burgess* insiste sur les obligations de l’expert, « à savoir l’impartialité, l’indépendance et l’absence de parti pris » de son opinion, c’est-à-dire que l’opinion de l’expert doit être objectif, dénué de toute influence externe et dénué

---

[56] *R c Abbey*, [1982] 2 RCS 24, 1982 CanLII 25 à la p 42 [*Abbey RSC*] en citant l’arrêt Turner de 1974.

[57] *R c Morin*, [1988] 2 RCS 345, 1988 CanLII 8 .

[58] *Mohan*, *supra* note 26 à la section III1c.

[59] *Ibid* à la section III1d.

de tout favoritisme d'une partie aux dépens de l'autre[60]. En somme, l'opinion de l'expert ne devrait pas changer, peu importe la partie qui a retenu ses services.

Si le témoignage ne satisfait pas aux quatre critères ci-haut, il ne devrait pas être admis[61]. Si les critères sont respectés, l'analyse se poursuit avec la deuxième étape qui inclut le critère de fiabilité.

1.2.2 Deuxième étape : l'analyse des avantages versus les risques en cas d'inclusion de la preuve

Cette étape constitue l'analyse du risque d'admission de la preuve par rapport à ses avantages. Dit autrement, la *valeur probante* d'un témoignage d'expert doit l'emporter sur ses effets préjudiciables. La *valeur probante* représente la force de conviction ou la capacité à rapprocher le tribunal de la vérité, impliquant que la preuve soit objective et de qualité, donc non viciée par l'erreur[62]. Les effets préjudiciables peuvent inclure les délais, les coûts ou la confusion causée par le témoignage auprès des juges des faits.

1.2.2.1 La valeur probante et les effets préjudiciables de la preuve

Ainsi l'arrêt *Mohan* considère que la preuve peut être exclue « si sa valeur n'en vaut pas le coût », en l'occurrence si les délais qu'elle occasionne sont déraisonnables : « si elle exige un temps excessivement long qui est sans commune mesure avec sa

---

[60] *White Burgess*, *supra* note 45 au para 28.

[61] *Bingley*, *supra* note 34 au para 14.

[62] André Émond, *Introduction au droit canadien,* 2e éd, Montréal, Wilson & Lafleur, 2016 à la p 82.

valeur ou si elle peut induire en erreur en ce sens que son effet sur le juge des faits, en particulier le jury, est disproportionné par rapport à sa fiabilité »[63].

Les « inconvénients » à prendre en considération incluent notamment l'influence indue que pourrait avoir une telle preuve auprès des juges des faits :

> La preuve d'expert risque d'être utilisée à mauvais escient et de fausser le processus de recherche des faits. Exprimée en des termes scientifiques que le jury ne comprend pas bien et présentée par un témoin aux qualifications impressionnantes, cette preuve est susceptible d'être considérée par le jury comme étant pratiquement infaillible et comme ayant plus de poids qu'elle ne le mérite.[64]

Ainsi, le juge des faits peut être confondu par la « mystique des sciences », si la preuve est susceptible de l'embrouiller, de le dérouter ou de l'écraser[65].

En somme, la preuve ne doit pas aboutir en une confiance aveugle des jurés envers l'expert, ce qui pourrait notamment avoir pour effet potentiel de détruire la présomption d'innocence et ce d'autant plus si la technique est complexe[66].

---

[63] *Mohan*, *supra* note 26 à la section III1a.

[64] *R c Trochym*, 2007 CSC 6, [2007] 1 RCS 239 au para 24 [*Trochym*].

[65] *R v Melaragni*, 1992 CanLII 12764 (ON SC) à la p 353: « is [the evidence] likely to confuse and confound the jury? (2) Is the jury likely to be overwhelmed by the « mystic infallibility » of the evidence ». La traduction des termes « confuse, confound, overwhelm » en « embrouiller, dérouter, écraser » provient de *Mohan*, *supra* note 26.

[66] Grimaud, *supra* note 20.

1.2.2.2 Le pouvoir discrétionnaire des juges

Puisqu'il s'agit ici de « préserver le procès devant juge et jury, et non pas d'y substituer le procès instruit par des experts »[67], dès lors, à cette étape, le juge du procès « conserve le pouvoir discrétionnaire d'exclure un témoignage qui satisfait aux critères minimaux d'admissibilité si les risques de son admission l'emportent sur ses avantages »[68].

Pour clore cette section, notons que l'admissibilité d'une preuve scientifique peut varier dans le temps, tenant ainsi compte de l'évolution des sciences : « Tout comme la communauté juridique, la communauté scientifique remet en question et améliore sans cesse ses connaissances fondamentales. L'admissibilité de la preuve scientifique n'est pas figée dans le temps »[69].

1.2.3 La fiabilité de la preuve scientifique

1.2.3.1 Le lien entre la fiabilité et la valeur probante de la preuve

Comme nous l'avons vu plus tôt, la Cour suprême dans l'arrêt *Mohan* prévoit que « la preuve d'expert qui avance une nouvelle théorie ou technique scientifique est soigneusement examinée pour déterminer si elle satisfait à la norme de fiabilité »[70].

---

[67] *White Burgess*, *supra* note 45 au para 18.

[68] *Bingley*, *supra* note 34 au para 16.

[69] *Trochym*, *supra* note 65 au para 31.

[70] *Mohan*, *supra* note 26 à la p 25; *J.-L.J.*, *supra* note 30.

La fiabilité peut être considérée comme partie intégrante de la deuxième étape de l'évaluation de l'admissibilité d'une preuve, soit celle de l'analyse des avantages et inconvénients de l'admission de la preuve d'expert. En effet, cette analyse considère la valeur probante de la preuve, laquelle implique que la preuve soit « suffisamment fiable pour franchir le seuil de l'admissibilité »[71].

L'arrêt *White Burgess* ajoute explicitement le critère de fiabilité à la démarche d'admissibilité de la preuve[72]. Ce dernier arrêt énonce que « dans le cas d'une opinion fondée sur une science nouvelle ou contestée ou sur une science utilisée à des fins nouvelles, la *fiabilité* des principes scientifiques étayant la preuve doit être démontrée » [nos italiques][73].

Une science « nouvelle » est définie par la Cour suprême dans l'affaire *Trochym*. La Cour affirme qu'« une technique ou connaissance scientifique sera considérée « nouvelle » dans deux cas : quand elle est nouvelle ou quand, bien qu'elle soit déjà reconnue, son application est nouvelle »[74].

---

[71] Paciocco, *supra* note 46 à la section 4.3(e)(i); *Manuel scientifique*, *supra* note 39 à la p 191.

[72] Paciocco, *supra* note 46 à la section 4.3(e)(i); *Manuel scientifique*, *supra* note 39 à la p 21.

[73] *White Burgess*, *supra* note 45 au para 23.

[74] *Trochym*, *supra* note 65 au para 133.

1.2.3.2 La fiabilité en droit

Il convient de préciser que la définition du terme « fiabilité » tel qu'entendue en droit correspond à ce que les scientifiques appellent plutôt la « *validité scientifique* ». Le professeur Paciocco (avant qu'il ne devienne juge) indiquait :

> Les scientifiques font une distinction entre la « validité » (le principe en cause démontre-t-il ce qu'il doit démontrer?) et la « fiabilité » (l'application du principe en cause donne-t-elle des résultats uniformes?). Les tribunaux qui traitent de la fiabilité dans ce contexte veulent parler de la fiabilité de la preuve, ou de ce que les scientifiques appelleraient la « validité scientifique »[75].

La « validité scientifique » porte sur le contrôle qu'un résultat soit obtenu selon la méthode scientifique et non le fruit du hasard ou dû à une lacune des appareils de mesures, de l'environnement, de l'échantillon ou du processus de production du résultat[76]. La fiabilité inclut donc l'ensemble des méthodes utilisées pour produire le résultat présenté en preuve.

1.2.3.3 Les critères d'évaluation de la fiabilité

Lorsque la preuve est d'ordre scientifique, les critères de fiabilité provenant de la décision *Daubert c. Merrell Dow Pharmaceuticals, Inc*[77] aux États-Unis sont

---

[75] Paciocco, *supra* note 29.

[76] Fiona Fidler et John Wilcox, « Reproducibility of Scientific Results » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, summer 2021 éd, Metaphysics Research Lab, Stanford University, 2021. La section 1.2 traite de l'impact de divers paramètres de test sur la validité des résultats, selon Gómez, Juristo et Vega.

[77] *Daubert v Merrell Dow Pharmaceuticals Inc*, 509 US 579, 1993 US Lexis 4408 [*Daubert*].

appliqués[78]. Bien que la Cour suprême du Canada rappelle dans l'arrêt *J.-L.J.* que *Daubert* se doit d'être interprété selon les *Federal Rules of Evidence*[79] qui diffèrent des règles de procédures du Canada, elle reprend explicitement les quatre critères énumérés dans cette affaire américaine afin d'évaluer la fiabilité d'une nouvelle technique scientifique[80] :

1- **La revue par les pairs** : La théorie ou la technique a-t-elle fait l'objet d'un contrôle par des pairs et d'une publication?
2- **Le taux d'erreur** : Le taux connu ou potentiel d'erreur ou l'existence de normes
3- **Les tests** : La théorie ou la technique peut-elle être vérifiée et l'a-t-elle été?
4- **L'acceptation générale** : La théorie ou la technique utilisée est-elle généralement acceptée?

Ces critères ne doivent pas être appliqués à la lettre, ils dépendent du contexte[81]. Ils se doivent d'être flexibles et non-exclusifs[82]. Dans l'affaire *R. c. Abbey*, le professeur Paciocco donne en exemple les théories actuarielles basées sur des méthodes

---

[78] Pierre Patenaude, « De Mohan à J.-L.J., de Daubert à Khumo: qu'en est-il de la preuve scientifique ou technique innovatrice? » dans Développements récents en droit administratif et constitutionnel 2002, Cowansville, Yvon Blais, 2002:« Le juge Binnie, dans une décision unanime, eut recours à la grille d'analyse élaborée par la Cour suprême des États-Unis dans l'arrêt *Daubert c. Merrell Dow Pharmaceuticals Inc.*; on y avait énuméré certains facteurs susceptibles d'être utiles pour évaluer la solidité, la fiabilité d'une nouvelle théorie ou technique scientifique.». note 1 à la p 195.

[79] *Federal Rules of Evidence*, 28 USC (2022).

[80] *J.-L.J.*, *supra* note 30 au para 33. Nous avons modifié l'ordre pour la fluidité de l'analyse qui suivra.

[81] *Abbey RSC*, *supra* note 57 au para 114.

[82] *Trochym*, *supra* note 65 au para 139 : « Les facteurs énumérés dans *Daubert c. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993)*, repris dans *J.-L.J.*, se voulaient flexibles et non exclusifs ».

statistiques qui ne suivent pas la « méthode scientifique » à proprement parler mais ne sont pas inadmissibles pour autant :

> Clearly it is inappropriate to consider all expertise as science, or to require all expertise to attain the scientific method. Some expert witnesses rely on science only in a loose sense. Actuaries apply probability theory and mathematics to produce decidedly unscientific results[83].

Nous expliquons chacun des critères dans ce qui suit.

1.2.3.3.1 Critère #1 – La revue par les pairs

Ce critère fait appel à la communauté scientifique pour s'assurer que la théorie ou technique n'est pas viciée, que la méthode scientifique a bien été respectée. La revue par les pairs sert de garde-fou aux lacunes scientifiques et à l'utilisation erronée de la théorie ou technique :

> [L]'assujettissement à l'examen de la communauté scientifique fait partie de l'« application rigoureuse de la démarche scientifique », en partie parce qu'il augmente les chances de déceler des failles importantes dans la méthode en cause[84].

La revue par les pairs s'effectue habituellement par l'entremise de publications dans des revues scientifiques. C'est d'ailleurs « l'existence de publications spécialisées »[85] qui permet et fait foi d'un débat ouvert sur la question.

---

[83] *Abbey RSC*, *supra* note 57 au para 114.

[84] *J.-L.J.*, *supra* note 30 au para 33.

[85] Paciocco, *supra* note 29 à la p 448.

La juge Deschamps dans *Trochym* en appelle non seulement à la communauté scientifique mais également à la communauté juridique car il importe d'évaluer la théorie ou technique à l'aune du droit : « … les juristes ont beaucoup écrit sur l'hypnose. Comme notre analyse porte sur la fiabilité de cette technique dans le contexte judiciaire, ces sources nous seront utiles »[86]. La juge analyse donc l'état du débat juridique au sujet de l'hypnose, les consensus et les différends qui s'en dégagent.

En résumé, la technique doit être acceptée par la communauté scientifique et juridique. On veillera à relever les enjeux et les aspects consensuels ou non du débat sur la technique.

1.2.3.3.2 Critère #2 – Le taux d'erreurs

Tout résultat empirique comprend une certaine incertitude laquelle peut s'immiscer à n'importe quelle étape du processus de production du résultat[87]. Ici, tel que spécifié dans l'arrêt *J.-L.J*[88]*,* l'incertitude en question concerne le taux d'erreur[89]. On dira des résultats obtenus à l'aide de théorie ou technique scientifiques qu'ils sont ou non « statistiquement significatifs » selon des normes statistiques établies en science[90]. L'erreur peut être de deux natures : un faux positif ou un faux négatif. Dit

---

[86] *Trochym*, *supra* note 65 au para 39.

[87] Limpert, « Beyond the Rule in Mohan », *supra* note 27 à la p 84.

[88] *J.-L.J.*, *supra* note 30.

[89] P Brad Limpert, « Beyond the Rule in *Mohan*: A New Model for Assessing the Reliability of Scientific Evidence » (1996) 54:1 U Toronto Fac L Rev 65-106. Le taux d'erreur n'est qu'une facette de l'incertitude selon l'auteur. L'auteur définit un cadre qui inclut plusieurs types d'incertitude, tels que l'incertitude concernant la modélisation, la communication, etc.

[90] *Manuel scientifique*, *supra* note 39 à la p 101.

autrement par le professeur Paciocco : « la nature des erreurs potentielles : sont-elles prédisposées à causer de fausses inclusions ou de fausses exclusions? »[91].

En résumé, les résultats empiriques se mesurent à l'aide de deux valeurs : le taux de fausses inclusions (faux positifs) ou le taux de fausses exclusions (faux négatifs). Nous définissons ces termes dans la section qui traite du taux d'erreurs[92].

1.2.3.3.3 Critère #3 – Les tests

Ce critère est celui de la vérification des hypothèses. En effet, la science exige que les affirmations (c'est-à-dire les hypothèses) soient testées[93]. Comme le soulignait la Cour suprême en empruntant à l'arrêt américain *Daubert*, « [l]a méthode scientifique actuelle est fondée sur la formulation d'hypothèses et leur vérification pour voir si elles sont fausses; en réalité, cette méthode est ce qui distingue la science des autres domaines de la connaissance »[94].

Bien que la vérification d'hypothèses soit effectivement la norme en sciences, elle peut ne pas s'appliquer en pratique lorsque par exemple le test nécessite des instruments d'une précision non disponible[95]. C'est ainsi que la Théorie de la relativité

---

[91] Paciocco, *supra* note 29 à la p 449.

[92] La section 7.2.1.1 traite du taux d'erreurs en termes de faux positifs et faux négatifs.

[93] Paciocco, *supra* note 29 à la p 448. Le critère de test est exprimé ainsi par le Professeur Paciocco : « si la théorie ou la technique découle de l'application d'une "méthode scientifique" et si elle a été "validée convenablement".

[94] *J.-L.J.*, *supra* note 30 au para 33.

[95] *Manuel scientifique*, *supra* note 39 à la p 57.

d'Einstein était vérifiable en théorie mais n'a pu être vérifié en pratique que des années plus tard grâce aux avancées dans la précision de nos instruments[96].

Les tests scientifiques sont généralement effectués sur des échantillons, c'est-à-dire un sous-ensemble de données qui doit être représentatif de la réalité. Ce point a été soulevé dans l'affaire *Ewert c. Canada*[97] où la fiabilité d'un outil de profilage psychologique utilisé par le Service correctionnel du Canada a été contestée car aucun test n'avait été effectué sur les autochtones, communauté dont faisait partie l'inculpé. La Cour a reconnu l'obligation du Service correctionnel de corriger cette lacune :

> s'il veut continuer à se servir des outils contestés, il doit à tout le moins mener des recherches pour savoir si, et le cas échéant dans quelle mesure, ces outils sont susceptibles de donner lieu à de la variance interculturelle lorsqu'on les utilise à l'égard de délinquants autochtones[98].

On doit également veiller à ce que l'ensemble des conditions de tests soient représentatives de la réalité. Une grande variabilité entre les conditions de tests et la réalité pourra fausser les résultats de tests. Ce point a été souligné dans *Trochym* où il est question de la technique d'hypnose qui y a été contestée pour l'effet des émotions sur les résultats de tests :

> Les études faites en laboratoire sont généralement abstraites et dépourvues de l'aspect émotionnel ou du sens qui se rattachent normalement aux souvenirs « réels » … Il se pourrait donc que les conclusions auxquelles

---

[96] *Ibid*.

[97] *Ewert c Canada*, [2018] 2018 CSC 30, [2018] 2 RCS 165 [*Ewert*].

[98] *Ibid* au para 67.

> elles arrivent sur l'hypnose ne s'appliquent guère dans le domaine de la criminalistique[99].

Notons finalement deux caractéristiques des tests scientifiques : la reproductibilité et la réplicabilité. La reproductibilité réfère à la capacité d'un chercheur d'obtenir les mêmes résultats lorsqu'un test est effectué à nouveau dans les mêmes conditions[100]. Les *techniques* (par opposition aux *sciences*) sont particulièrement caractérisées par une grande reproductibilité[101]. Les tests doivent également pouvoir être *répliqués*, c'est-à-dire, le test doit pouvoir être répété avec des données différentes et produire les mêmes résultats[102]. Plus un test est répliqué moins il y a de risque de considérer l'hypothèse testée vraie alors qu'elle est fausse. L'affaire *R v Abbey*[103] jugée par la Cour d'appel de l'Ontario réfère directement à la réplicabilité des tests (*replicate results*) : « Those factors [*Daubert*] include the existence of measurable error rates, peer review of results, the use of random sampling and the ability of the tester to replicate his or her results »[104].

---

[99] *Trochym*, *supra* note 65 au para 38.

[100] Kenneth Bollen et al, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*, Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, National Science Foundation, 2015 : « Reproducibility is a minimum necessary condition for a finding to be believable and informative ».

[101] *Manuel scientifique*, *supra* note 39 à la p 60.

[102] Bollen et al, *supra* note 101.

[103] *R v Abbey*, [2009] 2009 ONCA 624, 97 OR (3e) 330 [*Abbey ONCA*].

[104] *Ibid* au para 104.

En résumé, lorsque possible, les tests doivent être effectués sur des échantillons représentatifs de la réalité et dans des conditions représentatives de la réalité. Ils doivent être réplicables (obtenir les mêmes résultats avec des données distinctes) et reproductibles (obtenir les mêmes résultats avec les mêmes données).

1.2.3.3.4 Critère #4 – L'acceptation générale

Au sujet de l'acceptation générale par la communauté scientifique, l'arrêt *J.-L.J.* mentionne qu' « une technique connue qui n'a obtenu qu'un appui minimal au sein de la communauté, [. . .] peut à juste titre être envisagée avec scepticisme »[105].

La juge Deschamps dans *Trochym* se réfère à l'acceptation scientifique *dans le cadre juridique* : « [c]omme je l'ai mentionné, les opinions sont partagées au sein de la communauté scientifique quant à savoir si l'utilisation de l'hypnose à des fins judiciaires est acceptable »[106]. La juge se réfère abondamment à la jurisprudence internationale (Royaume-Uni, Australie, Nouvelle-Zélande, États-Unis). Dans cette affaire, les juges dissidents relèvent qu'il ne doit pas nécessairement y avoir unanimité au sein de la communauté scientifique. En effet, « [c]e critère ne précise pas quelle proportion d'experts correspond à l'acceptation générale. Les tribunaux n'ont jamais exigé l'unanimité, et tout ce qui ne fait pas consensus absolu en science peut rapidement prendre la forme d'un désaccord profond »[107].

---

[105] *J.-L.J.*, *supra* note 30 au para 33.

[106] *Trochym*, *supra* note 65 au para 47.

[107] *Ibid* au para 140.

1.2.3.4 Les aspects de la fiabilité

1.2.3.4.1 La fiabilité du principe fondamental

Le principe fondamental ou l'assise scientifique d'une nouvelle technique ou science doit faire l'objet d'un examen soigneux tel que précisé par la Cour suprême dans *R. c. Bingley* [108]. Ainsi, avant même d'examiner les méthodes utilisées pour produire un résultat, le professeur Patenaude suggère de commencer par vérifier la validité du principe fondamental de la technique[109], ce qui pourra dans certains cas rendre irrecevable un résultat si ce dernier est construit sur un principe fondamental qui n'est pas établi[110].

En effet, lorsque les techniques ne sont pas établies par une procédure judiciaire, ses principes ne peuvent être présumés. Toujours dans *R. c. Bingley*, la Cour suprême ajoute :

> Pour pouvoir admettre en preuve une opinion d'expert fondée sur des techniques ou des connaissances scientifiques, le juge de première instance doit avoir l'assurance que les principes scientifiques sous-jacents sont suffisamment fiables. Lorsque le recours à ces principes dans des procédures judiciaires est bien établi, le juge disposera souvent de précédents où des éléments de preuve fondés sur ces principes ont déjà été déclarés admissibles… En revanche, *le degré minimal de fiabilité requis d'un témoignage fondé sur des principes scientifiques nouveaux doit être établi dans le cadre d'un voir-dire, parce que la fiabilité et la validité des*

---

[108] *Bingley*, *supra* note 34 au para 41.

[109] *Hotel central (Victoriaville) inc Ltée c Compagnie d'assurance reliance Ltée*, [1998] 1998 CanLII 12934 (QC CA) n 22 [*Hotel Central*].

[110] Pierre Patenaude, « De l'expertise judiciaire dans le cadre du procès criminel et de la recherche de la vérité : quelques réflexions » (1997) 27:1 RDUS 1-47 à la p 20.

> *prémisses sur lesquels ceux-ci reposent ne sauraient être présumées* [nos italiques][111].

Et la Cour complète en affirmant que

> [s]'il est incapable de vérifier la fiabilité du fondement scientifique de l'évaluation […], le juge du procès — dans son rôle de gardien du processus judiciaire — ne sera pas en mesure d'apprécier la valeur probante d'une telle preuve, et le juge des faits sera incapable de déterminer le poids à y accorder[112].

1.2.3.4.2 La fiabilité de la méthode

La méthode utilisée pour produire le résultat devra respecter la méthode scientifique. Le professeur Patenaude souligne que pour ceux qui ont à interroger les experts, la règle d'or est de comprendre que la valeur de leur témoignage dépend principalement de la validité des méthodes utilisées[113]. Par exemple, bien que l'analyse par ADN soit couramment acceptée en preuve sur la base de sa fiabilité[114], il arrive que les procédures des laboratoires en charge de produire une analyse comportent de nombreuses lacunes, ce qui incitera les juristes à s'imposer un devoir de doute judicieux quant aux méthodes utilisées[115].

---

[111] *Bingley*, *supra* note 34 au para 43.

[112] *Ibid* au para 40.

[113] Patenaude, Pierre, *supra* note 27 à la p 412.

[114] Jill R Presser et Kate Robertson, *AI Case Study: Probabilistic Genotyping DNA Tools in Canadian Criminal Courts*, Commission du droit de l'Ontario, 2021 à la p 10.

[115] Patenaude, Pierre, *supra* note 27 aux pp 410, 411.

1.2.3.4.3 La fiabilité de l'explication de la preuve

En common law, la tradition veut que la preuve puisse être expliquée[116]. L'explication de la preuve contribue à la légitimité de la décision finale :

> des motivations qui contiennent des explications détaillées et claires de la preuve elle-même, ainsi que du raisonnement du jugement lui permettant de distinguer entre une preuve plus faible et plus forte, confèrent un sens de légitimité à la décision finale[117].

Le rôle de l'expert est justement de fournir une explication permettant « l'éclairage du juge sur des questions scientifiques ou techniques d'une certaine complexité, dépassant l'expérience ou les connaissances de ce dernier »[118]. L'expert se pose comme l'interprète des résultats obtenus par le processus scientifique, tel que le rappelle *R. c. Béland* :

> [T]oute scientifique que puisse être cette preuve, son utilisation devant le tribunal dépend d'une intervention humaine. Quels que soient les résultats […], c'est par la bouche de l'expert que leur nature et leur sens sont communiqués au juge des faits. La faillibilité humaine est par conséquent toujours présente[119].

Selon Limpert l'explication est liée à la fiabilité de la preuve elle-même car elle touche à l'incertitude de la preuve. Dans ce contexte, Limpert définit l'incertitude

---

[116] Patrick Nutter, « Machine Learning Evidence: Admissibility and Weight » (2019) 21:3 U Pa J Const L 919 à la p 947 : « common law tradition has always required evidence that is explainable, bears discernible logic, and may be examined or challenged ».

[117] Stettler, *supra* note 31 au para 91. L'auteur cite la juge canadienne J. Smith.

[118] *Ibid* au para 30.

[119] *R. v. Béland*, *supra* note 35 au para 20.

comme étant le degré de précision ou la validité avec laquelle l'explication rend compte d'un phénomène observable[120]. L'incertitude et la fiabilité de la preuve sont liées, l'incertitude de la preuve scientifique permet d'établir le niveau de fiabilité de la preuve[121].

Le professeur Paciocco s'en réfère à la « clarté avec laquelle la technique [sous-jacente à la preuve scientifique] peut être expliquée »[122]. L'explication a trait au « caractère appréciable de la preuve », c'est-à-dire si est-elle susceptible d'aider le jury ou au contraire de l'embrouiller[123]. Lors du voir-dire[124], l'expert doit être en mesure d'expliquer le résultat spécifique produit, c'est-à-dire l'expert doit justifier comment la technique a produit le résultat en question plutôt qu'un autre[125]. Dans quelle mesure

---

[120] Limpert, « Beyond the Rule in Mohan », *supra* note 27 n 130.

[121] *Ibid* à la p 84.

[122] Paciocco, *supra* note 24 à la p 450. L'auteur fait référence à l'arrêt *R. v. Johnston* à la note 129. *R v Johnston*, 1992 Superior Court of Justice of Ontario.

[123] David Paciocco, « L'évaluation du témoignage d'opinion pour en établir l'admissibilité : les leçons récentes du droit de la preuve » (1995) 26:3 Revue générale de droit 425-454 à la p 450. L'auteur fait référence à *R. c. Mohan* à la note 128. *R c Mohan*, [1994] 2 RCS 9.

[124] Nicolas Bellemare, « Chapitre III - Les procédures précédant le procès en matière criminelle » dans *Droit pénal: procédure et preuve* Collection de droit 2019-2020, Montréal, Québec, Éditions Yvon Blais, 2019, 41 à la p 83. Le « voir-dire » est la procédure qui permet au juge du procès de statuer sur l'admissibilité de la preuve en interrogeant le témoin.

[125] *R. v. Béland*, *supra* note 35. La Cour souligne que le résultat produit par une technique, en l'occurrence le détecteur de mensonge, n'est accessible à la Cour qu'au travers les explications d'un expert et met en garde contre l'interprétation de l'expert : « il faut se rappeler que toute scientifique que puisse être cette preuve, son utilisation devant le tribunal dépend d'une intervention humaine, celle de l'expert en détecteurs de mensonges. Quels que soient les résultats enregistrés par le détecteur de mensonges, c'est par la bouche de l'expert que leur nature et leur sens sont communiqués au juge des faits ». Stettler, *supra* note 31 : « Les experts doivent [...] être en mesure d'expliquer le comment et le pourquoi de leurs conclusions ».

l'expert peut-il expliquer le résultat, quel est le niveau d'interprétation requis et dans quelle mesure le résultat peut-il être contre-validé? Le Tribunal doit pouvoir jauger ces questions afin d'accorder la juste valeur probante au témoignage[126]. L'explication revêt un intérêt particulier dans le cas de l'apprentissage automatique qui, comme nous le verrons dans les prochains chapitres, pose des enjeux d'intelligibilité dûs à la complexité de la technique.

1.3 Résumé de la démarche d'admissibilité de la preuve scientifique

La preuve scientifique est admissible en cour en tant que témoignage d'expert selon la démarche suivante :

- Étape 1 : la preuve doit respecter les quatre critères suivants :
    - La pertinence : elle permet à la Cour de se rapprocher de la vérité;
    - La nécessité d'aider le juge des faits : le juge des faits n'a pas les connaissances voulues pour apprécier la preuve et se doit de faire appel à un expert;
    - L'absence de toute règle d'exclusion : la preuve ne contrevient à aucune autre règle de droit;
    - L'expertise particulière : l'expert possède des connaissances spécifiques.
- Étape 2 : si les critères de l'étape 1 sont rencontrés, les avantages et les effets préjudiciables pouvant découler de l'admission de la preuve sont analysés. La valeur probante de la preuve doit être proportionnée aux risques encourus par son admission, lesquels peuvent inclure des délais, des coûts et de la confusion auprès de la Cour.
    - La valeur probante est fonction de la *fiabilité* de la preuve. Dans le cas d'une nouvelle technique ou science, la fiabilité ou validité scientifique

---

[126] Roth, *supra* note 23. L'article concerne les enjeux d'opacité « black box dangers » des techniques utilisées en preuve.

de la preuve doit être soigneusement examinée. La fiabilité de la preuve inclut :

- le principe fondamental sur lequel repose la nouvelle technique ou science,
- l'ensemble des méthodes utilisées pour produire un résultat,
- l'explication de la preuve, explication qui par sa clarté, validité et précision contribue à la fiabilité de la preuve.

- La fiabilité de la preuve scientifique est évaluée selon les critères suivants :
  - La revue par les pairs, c'est-à-dire les enjeux et conclusions émanant d'un débat public et transparent auprès des communautés scientifique et juridique à propos de la nouvelle technique ou science;
  - Les tests effectués pour obtenir les résultats présentés en preuve;
  - L'incertitude ou le taux d'erreurs associé à la preuve;
  - L'acceptation générale de la preuve dans un cadre juridique.

Le prochain chapitre porte sur l'apprentissage automatique, ses composantes et son fonctionnement.

CHAPITRE II

L'APPRENTISSAGE AUTOMATIQUE

Dans ce chapitre nous décrivons l'apprentissage automatique, ses principales caractéristiques ainsi que le processus de développement d'un outil d'apprentissage automatique.

2.1 Qu'est-ce que l'apprentissage automatique

L'apprentissage automatique est une spécialisation de l'intelligence artificielle[127]. Nous ferons donc un bref détour par l'intelligence artificielle (IA) afin de mieux situer l'apprentissage automatique. « Can machines think? ». Cette question lancée par Alan Turing[128] en 1950 et qui obsède les penseurs depuis des siècles, est au cœur des préoccupations de « l'intelligence artificielle », terme consacré en 1956[129].

---

[127] Bringsjord et Govindarajulu, *supra* note 24 à la section 4.

[128] Andrew Hodges, « Alan Turing » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2019 éd, Metaphysics Research Lab, Stanford University, 2019. Alan Turing (1912-1954) est un mathématicien anglais dont les travaux ont largement contribué à la naissance de l'intelligence artificielle.

[129] Bringsjord et Govindarajulu, *supra* note 24 à la section 1. On peut tisser des liens étroits entre l'IA et la philosophie car les deux champs se penchent sur la mécanique de la pensée rationnelle. Aristote avec

Russel et Norvig, dans un ouvrage reconnu comme une autorité, « *Artificial Intelligence : A Modern Approach* »[130], définissent le champ de l'intelligence artificielle selon les objectifs qu'il tente d'accomplir. Ceux-ci sont au nombre de quatre[131], soit :

- le comportement humain;
- la pensée humaine;
- le comportement rationnel idéal;
- la pensée rationnelle idéale.

On distingue deux grandes approches pour aborder ces objectifs, soit l'approche symbolique et l'approche connexionniste[132]. La première approche utilise le plus souvent des règles prédéfinies; c'est le cas par exemple des systèmes experts qui utilisent des structures en forme d'arbres de décisions pour parvenir à l'objectif[133].

---

sa théorie des syllogismes jette les assises de la pensée rationnelle; Descartes se pose la même question que Turing quatre siècles auparavant.

[130] Stuart Russell et Peter Norvig, *Intelligence artificielle: Avec plus de 500 exercices*, Pearson Education France, 2010; « Artificial Intelligence: A Modern Approach, 4th US ed. », en ligne: *Berkeley* <http://aima.cs.berkeley.edu/>.

[131] Bringsjord et Govindarajulu, *supra* note 24. Les quatre catégories ne sont pas nécessairement exhaustives ou mutuellement exclusives.

[132] Russell et Norvig, *supra* note 131.

[133] Josef Bajada, « Symbolic vs Connectionist A.I. », (9 avril 2019), en ligne: *Towards Data Science* <https://towardsdatascience.com/symbolic-vs-connectionist-a-i-8cf6b656927>.

Dans la seconde approche, l'approche connexionniste, les règles ne sont pas définies d'avance[134]. C'est l'approche utilisée par l'apprentissage automatique[135]. Elle peut être caractérisée par une structure qui tente d'imiter le cerveau humain, appelée « réseau de neurones »[136], d'où le terme *connexionnisme*[137]. Notre projet s'inscrit dans la branche de l'intelligence artificielle dont l'objectif est de calquer le comportement rationnel idéal selon une approche connexionniste. C'est la branche de l'apprentissage automatique (*machine learning*).

L'apprentissage automatique peut se définir ainsi :

> Machine learning is concerned with building systems that improve their performance on a task when given examples of ideal performance on the task, or improve their performance with repeated experience on the task[138].

Ainsi, le système « apprend » à accomplir une tâche, une approche qui se démarque de l'informatique classique où c'est plutôt l'humain qui fournit les instructions précises permettant d'accomplir une tâche[139].

---

[134] *Ibid*.

[135] Bringsjord et Govindarajulu, *supra* note 24. L'apprentissage automatique est un exemple du formalisme appelé connexionniste ou « réseau de neurones ».

[136] Copeland, *supra* note 1 à la section « Methods and Goals in AI » .

[137] Bajada, *supra* note 134. Bien qu'il existe différentes structures, les réseaux de neurones sont très répandus.

[138] Bringsjord et Govindarajulu, *supra* note 24.

[139] *Ibid*.

Pour apprécier le reste de la discussion, il convient de définir le terme *algorithme*. Un algorithme réfère intuitivement à une série d'instructions permettant d'accomplir une tâche[140]. Plus précisément, Markov[141] définit un algorithme comme suit : « a computational process that is *determined*, *applicable*, and *effective* » [nos italiques][142]. Il est déterminé (*determined*) car il implique que les instructions soient suffisamment précises pour ne laisser aucune place à un choix arbitraire. Il est généralisable (*applicable)* car les algorithmes s'appliquent à des ensembles de données et non une seule donnée précise, par exemple l'ensemble des nombres réels et non pas un seul nombre réel. Il est effectif (*effective)* car il produit une réponse à un problème donné. Un *programme* est l'implantation d'un algorithme dans un langage qu'une machine peut décoder[143]. La Figure 1[144] illustre la distinction entre la programmation classique et l'apprentissage automatique. Dans l'approche classique, un programme est fourni à la machine; c'est elle qui produit un résultat en exécutant le programme tandis que dans une approche par apprentissage automatique, c'est la machine qui produit le programme à partir des résultats fournis. Dans l'approche classique, l'humain conçoit

---

[140] Nicola Angius, Giuseppe Primiero et Raymond Turner, « The Philosophy of Computer Science » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2021 éd, Metaphysics Research Lab, Stanford University, 2021 à la section 3 : « Intuitively, an algorithm is a set of instructions allowing the fulfillment of a given task ».

[141] *Ibid*. Markov est un mathématicien, auteur de l'ouvrage « Theory of Algorithms », 1954.

[142] *Ibid* à la section 3.1 « Classical Approaches ».

[143] *Ibid* à la section 4 : « programs are implementations of algorithms ».

[144] Eric Grimson, John Guttag et Ana Bell, « Introduction to Computational Thinking and Data Science - MIT Course Number 6.0002 », (Automne 2016), en ligne: *MIT OpenCourseWare* <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/>. L'image a été extraite à 00h:08m:31s de la session 11 « Introduction to Machine Learning ».

un programme qui, exécuté par la machine, produira un résultat. Dans l'approche par apprentissage automatique, le programme[145] est conçu par la machine à partir des résultats fournis.

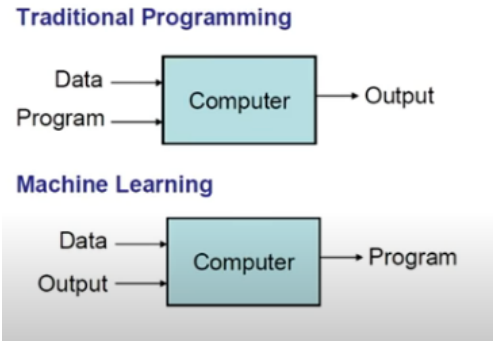

Figure 1 - Programmation traditionnelle et apprentissage automatique[146]

[145] Nous abordons plus à fond les composantes d'un système d'apprentissage automatique à la section 2.1.2.2.

[146] La figure est tirée de Grimson, Guttag et Bell, *supra* note 145 à la session 11, vers 00:08:00. Pour les fins de discussions, la figure est simplifiée. Les personnes intéressées peuvent se référer à une image plus complète ici: Sidharth Mehra, *Detection of Offensive Language in Social Media Posts*, mémoire de maîtrise en intelligence artificielle au département d'informatique, Cork Institute of Technology, 2020 [non publiée] à la figure 2.2. Pour une explication plus détaillée, voir Jason Brownlee, « Difference Between Algorithm and Model in Machine Learning », (28 avril 2020), en ligne: *Machine Learning Mastery* <https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/>.

2.1.1 Les modes d'apprentissage

Il existe des dizaines d'algorithmes d'apprentissage automatique[147] lesquels sont basés sur la théorie des probabilités[148]. Chaque algorithme est adapté à différents types de problèmes et données[149]. La Figure 2 en illustre quelques-uns. On les catégorise généralement selon trois grandes modes d'apprentissage : l'apprentissage supervisé, non supervisé et l'apprentissage par renforcement[150].

---

[147] Stuart Russell et Peter Norvig, *Intelligence artificielle: Avec plus de 500 exercices*, Pearson Education France, 2010, Google-Books-ID: DWTlFWSGxJMC.

[148] Kevin P Murphy, *Machine learning - A Probabilistic Perspective*, Adaptive computation and machine learning series, Cambridge, MA, MIT Press, 2012 à la p 28.

[149] *Ibid* à la p 1.

[150] Hooman Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods » (2019) 6 Academic Pathology.

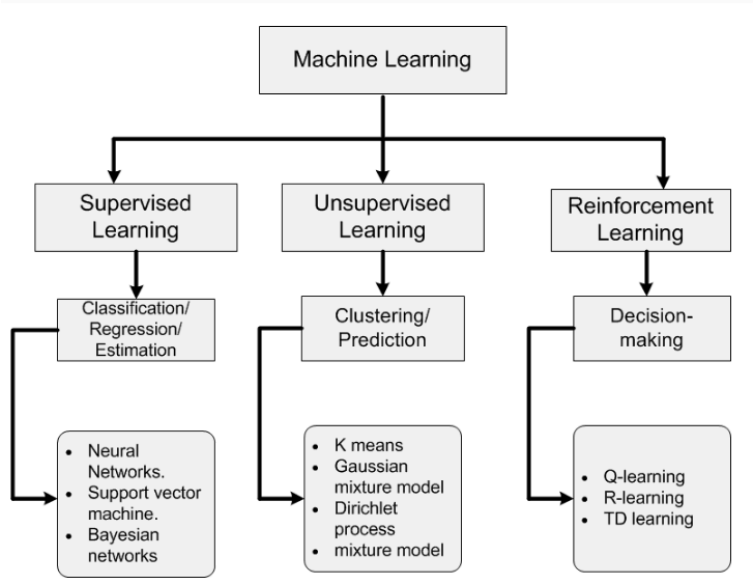

Figure 2 - Exemples d'algorithmes d'apprentissage automatique selon trois grandes approches d'apprentissage : supervisé, non-supervisé et par renforcement[151]

2.1.1.1 Apprentissage supervisé

Dans l'apprentissage supervisé, l'algorithme effectue une correspondance (*mapping*) entre des données d'entrée (*input*) et le résultat attendu en sortie (*output*)[152]. Le résultat attendu en sortie est fourni au système. Par exemple, pour qu'une machine apprenne à détecter une image de chat parmi une collection d'images diverses, on lui fournira des exemples d'images préalablement étiquetées « chat » et d'autres étiquetées « non-chat ». La correspondance entre les entrées et la sortie pourra par la suite être appliquée sur des images non étiquetées pour prédire la présence ou non d'un chat. La

---

[151] Grimson, Guttag et Bell, *supra* note 145. L'image a été tirée à 00h:09m:11s de la session 11 « Introduction to Machine Learning ».

[152] Russell et Norvig, *supra* note 131 à la p 737.

grande majorité des applications en intelligence artificielle utilise un mode d'apprentissage supervisé[153].

2.1.1.2 Apprentissage non supervisé

Dans l'apprentissage non supervisé, le système détecte des similitudes parmi des données non étiquetées, l'objectif étant de regrouper les données selon des similitudes qu'on ne connaît pas d'avance[154]. Cette approche est utilisée dans des problèmes de type exploratoires afin de déceler des structures insoupçonnées dans un grand volume de données[155], par exemple détecter des profils de consommateurs. On parlera de regroupements « agnostiques », c'est-à-dire sans idée préconçue des catégories[156].

2.1.1.3 Apprentissage par renforcement

Finalement, l'apprentissage par renforcement est un mélange des deux méthodes précédentes. Plutôt que de lui fournir le résultat attendu, l'algorithme obtient seulement un indice que le résultat est bon ou non[157]. L'algorithme apprend à partir d'interactions

---

[153] Bringsjord et Govindarajulu, *supra* note 24 à la section 4.1.

[154] Russell et Norvig, *supra* note 131 à la p 737.

[155] Bringsjord et Govindarajulu, *supra* note 24 à la section 4.1.

[156] Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology », *supra* note 151 à la p 8.

[157] M I Jordan et T M Mitchell, « Machine learning: Trends, perspectives, and prospects » (2015) 349:6245 Science 255.

avec l'utilisateur[158]. Un exemple est l'affichage publicitaire selon les préférences indiquées par le consommateur.

2.1.2 Les composantes d'un système d'apprentissage automatique

2.1.2.1 Les données

Les données constituent la matière première d'un système d'apprentissage automatique. Le processus d'apprentissage est entièrement dépendant des données fournies en entrée (*input*)[159]. Plus la quantité de données sera élevée, plus le système pourra converger vers le résultat désiré[160]. Certains systèmes sont particulièrement gourmands en données et nécessiteront des milliers, des millions, voire des milliards d'exemples pour donner des résultats exactes[161].

---

[158] Jonathan Schmidt et al, « Recent advances and applications of machine learning in solid-state materials science » (2019) 5:1 npj Computational Materials 1.

[159] EDUCBA, « Machine Learning Tutorial | Self Guides to Learn Machine Learning », (2020), en ligne: *EDUCBA* <https://www.educba.com/data-science/data-science-tutorials/machine-learning-tutorial/> à la section « Machine Learning Lifecyle » : « [...] model performance depends completely on the input data and the training process ».

[160] *Ibid* à la section « Machine Learning Lifecyle ».

[161] On parle essentiellement des systèmes par apprentissage profond, que nous abordons à la section suivante. Gary Marcus, « Deep Learning: A Critical Appraisal » (2018) arXiv:180100631 [cs, stat], en ligne: <http://arxiv.org/abs/1801.00631> à la p 7, arXiv: 1801.00631. Également à la page 5 : « In a world with infinite data, and infinite computational resources, there might be little need for any other technique ». Royaume-Uni, Centre for Data Ethics and Innovation (CDEI), « Facial Recognition Technology - Snapshot Paper », (28 mai 2020), en ligne: <https://www.gov.uk/government/publications/cdei-publishes-briefing-paper-on-facial-recognition-technology/snapshot-paper-facial-recognition-technology> à la p 12. Cette gourmandise en données pose des enjeux pour plusieurs organisations qui n'ont pas les ressources requises pour entraîner l'algorithme. Dans le domaine de la reconnaissance faciale par exemple : « Most organisations using trained FRT [Facial Recognition Technology] will not conduct the training themselves, and will instead purchase a pre-trained algorithm from a company with sufficient scale to do this work ».

2.1.2.2 Le modèle prédictif

Les algorithmes d'apprentissage automatique détectent donc des structures (*patterns*) dans les données. Lors de la phase d'apprentissage, ces structures seront raffinées jusqu'à l'obtention d'un modèle prédictif qui pourra par la suite être généralisé et appliqué à de nouvelles données[162]. Certains algorithmes nécessitent une sélection manuelle des caractéristiques (*features*) qui entreront dans la constitution du modèle[163]. On privilégiera celles dont les valeurs de prédiction sont les plus élevées. Par exemple, pour prédire l'obtention d'un diplôme universitaire, la moyenne générale d'un étudiant a une valeur prédictive plus élevée que la couleur de ses yeux[164]. D'autres algorithmes détectent automatiquement les caractéristiques prédictives à partir des données d'entrées; c'est le cas de l'apprentissage profond (*deep learning*)[165] utilisé dans des applications de reconnaissance faciale[166] et de reconnaissance de la parole[167], par exemple. Ces algorithmes, caractérisés par leurs multiples couches de neurones[168],

---

[162] Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology », *supra* note 151 à la p 5.

[163] Grimson, Guttag et Bell, *supra* note 145. On appelle « feature engineering » les activités d'optimisation des caractéristiques. Voir la vidéo à 00h:26m:00s.

[164] *Ibid*. L'exemple est tiré de la session 11 « Introduction to Machine Learning » à 00h:26m:00s.

[165] Alexander Amini, « Introduction to Deep Learning (6.S191) », (27 janvier 2020), en ligne: *MIT* <https://www.youtube.com/watch?v=njKP3FqW3Sk>, vers 08:00.

[166] Tom M Mitchell, « Does Machine Learning Really Work? » (1997) 18:3 AI Magazine 11 à la p 15.

[167] Jordan et Mitchell, « Machine learning », *supra* note 158.

[168] Eda Kavlakoglu, « AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference? », (27 juillet 2020), en ligne: *IBM* <https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks> : « It is the number of node layers, or depth, of neural networks that distinguishes a single neural network from a deep learning algorithm, which must have more than three ».

peuvent s'appliquer aux diverses approches d'apprentissage (supervisé, non supervisé ou renforcement), cependant ils sont le plus couramment utilisés en apprentissage supervisé[169].

2.1.2.3 Le résultat

Suite à l'apprentissage, lorsque le modèle prédictif est appliqué à des données, il produira un résultat. Selon le type de problème, le résultat pourrait être une catégorie (par exemple, une catégorie de 1 à 10 selon le niveau de risque de récidive d'un détenu ou une catégorie binaire oui/non), un classement (par exemple, le classement de documents selon leur pertinence), une valeur numérique (par exemple, le prix anticipé de vente d'une maison), une recommandation (par exemple, le film le plus susceptible de vous intéresser), ou une structure de données plus complexe (par exemple, une phrase en réponse à un courriel)[170].

2.1.3 Les étapes de production d'un système d'apprentissage automatique

Une fois entraîné sur un ensemble de données pour ajuster le modèle, ce dernier sera testé sur des données distinctes afin d'évaluer si sa performance est satisfaisante au point d'être déployé dans un contexte réel. Les étapes de développement d'un système d'apprentissage automatique sont donc les suivantes[171]:

---

[169] Ben Dickson, « Self-supervised learning: The plan to make deep learning data-efficient », (23 mars 2020), en ligne: *TechTalks* <https://bdtechtalks.com/2020/03/23/yann-lecun-self-supervised-learning/>.

[170] Les exemples sont tirés de ce site : Google, « Introduction to Machine Learning Problem Framing - Common ML Problems », (5 février 2021), en ligne: *Google Developers* <https://developers.google.com/machine-learning/problem-framing/cases?hl=fr>.

[171] EDUCBA, *supra* note 160 EDUCBA, supra note 151 à la section « Machine Learning Life Cycle ». Les étapes sont au nombre de huit. Nous les avons résumées. Pour plus de détails, voir David Lehr et

- Choisir l'algorithme d'apprentissage qui correspond le mieux aux objectifs du système;

- Préparer les données, incluant la collecte, la transformation, l'organisation et le nettoyage des données[172];

- Entraîner le système jusqu'à l'obtention d'un modèle;

- Tester le modèle sur de nouvelles données;

- Finalement, déployer le modèle dans un environnement réel. Les données réelles pourront être utilisées pour réentraîner le système afin de l'adapter continuellement à la réalité[173].

2.1.4 Exemples d'apprentissage automatique

Bien que l'apprentissage automatique existe depuis plusieurs décennies, la technique était auparavant inutilisable parce que trop lente; cependant, les avancées des procédés de fabrication des microprocesseurs ainsi les techniques de traitement de

---

Paul Ohm, « Playing with the Data: What Legal Scholars Should Learn about Machine Learning » (2017) 51:2 UCD L Rev 653 à la section II.

[172] Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology », *supra* note 151 à la p 4. Les auteurs définissent la préparation des données (*feature engineering*) comme suit : « This the process that allows for the selection of certain key features or the transformation/ conversion of certain features within a data set that will ultimately lead to a better prediction model ».

[173] Neoklis Polyzotis et al, « Data Management Challenges in Production Machine Learning » (2017) Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17) 1723-1726 fig 1. L'utilisation des données réelles pour réentraîner le système est illustrée à la figure 1 « Simplified schematic of a production machine learning pipeline ».

données massives ont aujourd'hui permis de combler cette lacune[174]. L'apprentissage automatique permet de résoudre des problèmes qui impliquent un certain degré d'incertitude[175]. Ainsi, la technique joue aujourd'hui un rôle essentiel dans notamment trois grandes familles d'applications[176] :

- Les applications prédictives basées sur des données historiques. Par exemple, un outil aidant à prédire le risque de complications dans le suivi médical de femmes enceintes[177];

- Les applications trop complexes pour être programmées de façon traditionnelle. Par exemple, un outil de reconnaissance de parole ou de reconnaissance faciale[178];

- Les applications personnalisées. Par exemple, un outil de suggestions de nouvelles (*newsfeed*) selon les préférences de l'utilisateur[179].

---

[174] Rick Swedloff, « The New Regulatory Imperative for Insurance » (2019) 61:6 Boston College L Rev 2033.

[175] Murphy, *supra* note 149 à la p 1.

[176] Mitchell, *supra* note 167.

[177] *Ibid*.

[178] *Ibid*.

[179] *Ibid*.

Au cours d'enquêtes policières, l'apprentissage automatique est utilisé dans des applications telles que la reconnaissance faciale[180], la reconstruction faciale par ADN[181], la reconnaissance de l'identité vocale[182], la reconnaissance d'armes à feu à partir de trace de balles[183], la reconnaissance de la marque de souliers laissée par une empreinte[184]. D'autres exemples d'applications utilisant l'intelligence artificielle incluent la surveillance auditive (par exemple, la surveillance de coups de feu), la lecture sur les lèvres, la détection de fraude fiscale et financement de terrorisme, les outils d'autopsie virtuelle pour aider à établir les causes de décès, la police prédictive, le suivi des médias sociaux et la détection des battements de cœur[185].

2.1.5 L'intelligence artificielle au Canada

2.1.5.1 Le secteur privé

Les données compilées par Global Advantage Consulting Group pour le compte de l'Université de Toronto révèlent qu'en 2020 le secteur privé de l'intelligence artificielle

---

[180] Lex Fridman, « Deep Learning Basics: Introduction and Overview (MIT course 6.S094) », (2019), en ligne: *Youtube* <https://www.youtube.com/watch?v=O5xeyoRL95U&list=PLrAXtmErZgOeiKm4sgNOknGvNjby9efdf>.

[181] Alicia Carriquiry et al, « Machine learning in forensic applications » (2019) 16 Significance (Royal Statistical Society) 29-35.

[182] *Ibid*.

[183] *Ibid*.

[184] *Ibid*.

[185] Ces exemples sont tirés de CE, *Résolution du Parlement européen du 6 octobre 2021 sur l'intelligence artificielle en droit pénal et son utilisation par les autorités policières et judiciaires dans les affaires pénales*, 2021.

au Canada était constitué de plus de 660 compagnies dont la majorité est localisée à Toronto et Montréal[186]. Des investissements de 3 milliards de dollars au cours de la dernière décennie ont permis au Canada de se hisser en première position au sein des pays du G7 et de la Chine en termes de nombre de brevets reliés à l'intelligence artificielle par million d'habitants et en quatrième position mondiale en termes d'environnement favorable à l'intelligence artificielle, derrière les États-Unis, la Chine et le Royaume-Uni[187].

2.1.5.2 Le secteur public

Dans une étude publiée en 2018, le groupe de recherche Citizen Lab de la faculté de droit de l'Université de Toronto rapporte que le Canada utilise ou compte utiliser des algorithmes prédictifs au sein de divers ministères[188]. Le rapport porte sur un outil en particulier utilisé par le ministère de l'Immigration pour accélérer le traitement des demandes d'immigration. L'étude rapporte également l'instigation d'un processus d'acquisition d'outils prédictifs utilisant l'intelligence artificielle pour le ministère du Revenu du Canada, le ministère de l'Emploi et du Développement social et le ministère de la Justice[189]. Finalement le rapport met en lumière les données

---

[186] Université de Toronto, *Canada's AI Ecosystem: Government Investment Propels Private Sector Growth,* 2020 [UT, « AI Ecosystem »]. Le chiffre représente le nombre de compagnies privées don l'activité principale est reliée à l'intelligence artificielle.

[187] Cette donnée est mesurée selon le Global AI Index qui recoupe une centaine d'indicateurs, voir *ibid*.

[188] Petra Molnar et Lex Gill, *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System*, International Human Rights Program, Faculty of Law, University of Toronto, 2018.

[189] *Ibid* à la p 15.

massives collectées et utilisées par le Service canadien du renseignement de sécurité pour fins de surveillance[190].

2.1.5.3 Les forces de l'ordre

Le rapport *To Surveil and Predict* publié en 2020 par le groupe de recherche Citizen Lab dresse un portrait des outils d'intelligence artificielle utilisés par les forces de l'ordre au Canada[191]. Le rapport recense les outils utilisés au Canada selon trois grandes familles, soit les outils de prédictions géographiques de crimes[192], les outils de prédictions de personnes susceptibles de commettre ou être victime de crimes[193] et les outils de surveillance tels que la surveillance des réseaux sociaux [194] ou la reconnaissance faciale. Le rapport de Citizen Lab conclut, quoiqu'avec certaines réserves, que l'utilisation d'outils d'intelligence artificielle est peu répandue parmi les forces de l'ordre au Canada[195], une caractéristique attribuable en partie à la prudence

---

[190] *Ibid* à la p 18.

[191] Kate Robertson, Cynthia Khoo et Yolanda Song, *To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada*, Citizen Lab, University of Toronto, 2020.

[192] *Ibid*. L'outil GeoDash est utilisé par les forces de l'ordre à Vancouver.

[193] *Ibid*. L'outil Gotham de la compagnie Palantir, par exemple, est utilisé par les forces policières de Calgary pour cerner les relations et comportements d'individus.

[194] *Ibid*, art 4.3. Les services de police de Toronto et Calgary ont fait usage de MediaSonar, par exemple.

[195] Robertson, Khoo et Song, *supra* note 192. En effet, l'utilisation de l'IA semble bien moins répandue au Canada qu'aux États-Unis, par exemple. Citizens Lab émet quelques réserves par rapport à ses résultats, réserves qui ont trait aux difficultés inhérentes à l'obtention de ce type d'information auprès des forces de l'ordre. Voir la section 4.4.

des législateurs et des forces de l'ordre elles-mêmes[196]. Notons à titre d'exemple l'outil de reconnaissance faciale ClearviewAI qui était utilisé par la GRC ainsi que les forces policières d'Edmonton, Toronto, Calgary, Vancouver, Halifax[197] avant d'être retiré du marché canadien en 2020 pour cause d'atteinte à vie privée suite à une enquête à laquelle participait le Commissariat à la protection de la vie privée du Canada[198].

2.2 Résumé

L'apprentissage automatique est une branche de l'intelligence artificielle. C'est une technique qui permet d'entraîner un système à partir des données pour qu'il formule un modèle qui lui permettra ensuite de produire un résultat voulu. Une fois entraîné et testé, le modèle résultant pourra être appliqué à de nouvelles données pour produire un résultat. Cette technique peut être utilisée par les forces de l'ordre, par exemple, au cours d'enquêtes policières pour identifier une personne, pour effectuer de la surveillance ou pour prédire des crimes. Le chapitre suivant plonge au cœur de la technique d'apprentissage automatique.

---

[196] *Ibid* à la section 4.5. Les contraintes budgétaires sont également identifiées comme cause probable de la relative absence d'outils.

[197] Céline Castets-Renard, Émilie Guiraud, et Jacinthe Avril-Gagnon, « Cadre juridique applicable à l'utilisation de la reconnaissance faciale par les forces de police dans l'espace public au Québec et au Canada. Éléments de comparaison avec les États-Unis et l'Europe » (2020) Observatoire international sur les impacts sociétaux de l'IA et du numérique, Chaire de recherche IA responsable à l'échelle mondiale, en ligne: <https://www.docdroid.com/YIDTjrr/cadre-juridique-applicable-a-lutilisation-de-la-reconnaissance-faciale-par-les-forces-de-police-dans-lespace-public-au-quebec-et-au-canada-pdf> à la p 12. La Sureté du Québec a pour sa part conclut une entente avec la société française Idemia pour l'achat d'un logiciel de reconnaissance faciale.

[198] Canada, Commissariat à la protection de la vie privée du Canada, « Clearview AI cesse d'offrir sa technologie de reconnaissance faciale au Canada », (6 juillet 2020), en ligne: <https://www.priv.gc.ca/fr/nouvelles-du-commissariat/nouvelles-et-annonces/2020/nr-c_200706/>.

CHAPITRE III

LA FIABILITÉ DU PRINCIPE FONDAMENTAL

3.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de son principe fondamental

Dans ce chapitre, nous soulevons les questions permettant d'évaluer la fiabilité du principe fondamental du système :

- Quel est l'objectif du système?
- Comment le système fonctionne-t-il globalement pour produire un résultat?
- Sur quelles hypothèses repose le modèle statistique?
- Que signifie le résultat, quelle est l'étendue des résultats possibles (par exemple, oui/non, une classe de 1 à 10)?
- Quelle est la classe de référence utilisée?
- En quoi le résultat est-il spécifique à la personne concernée?

3.2 L'approche statistique de l'apprentissage automatique

3.2.1 Déduction versus induction

L'apprentissage automatique procède par *induction* pour créer un modèle prédictif à partir d'exemples et par *déduction* lorsque ce modèle est appliqué à de nouvelles

données pour obtenir un résultat[199]. L'induction appelée aussi méthode par inférence statistique est la généralisation de cas particuliers : « Probabilistic inference consists in computing, from observed evidence expressed in terms of probability theory, posterior probabilities of propositions of interest »[200]. À l'inverse, la déduction procède du général (c'est-à-dire le modèle généré suite à la phase d'apprentissage) au particulier (c'est-à-dire le cas spécifique sous analyse par le modèle)[201].

Alors que la déduction donne lieu à une certitude quant à l'exactitude de l'application d'un modèle à un fait[202], l'induction repose sur des fondements plus « fragiles ou incertains »[203]. En effet, qu'est-ce qui peut justifier que l'on dérive une information future ou générale à partir d'observations sur des cas particuliers[204]? Les

---

[199] Jason Brownlee, « 14 Different Types of Learning in Machine Learning », (10 novembre 2019), en ligne: *Machine Learning Mastery* <https://machinelearningmastery.com/types-of-learning-in-machine-learning/>.

[200] Bringsjord et Govindarajulu, *supra* note 24 à la section 4.3.

[201] Brownlee, *supra* note 200.

[202] Bringsjord et Govindarajulu, *supra* note 24 : « The most significant difference between these forms of reasoning is that in the deductive case the truth of the premises guarantees the truth of the conclusion, whereas in the inductive case the truth of the premise lends support to the conclusion without giving absolute assurance ». Évidemment, il s'agit ici d'une certitude dans l'application et non dans la véracité du résultat; si la règle appliquée est erronée relativement au type de faits en cause, son application pourra être certaine, mais le résultat sera certainement erroné.

[203] Giovanni Busino, « La preuve dans les sciences sociales » (2003) 41:128 Revue européenne des sciences sociales 11 au para 7 : « Chez ces chercheurs l'induction (le passage par l'expérience immédiate et sensible pour aller du particulier au général) reste un point ferme, indubitable, en dépit du fait que sa légitimation et ses fondements demeurent fragiles, sinon incertains, que sa nature est fondamentalement heuristique ».

[204] Jan-Willem Romeijn, « Philosophy of Statistics » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2017 éd, Metaphysics Research Lab, Stanford University, 2017, à la section 1 : « There is no proper justification for procedures that take data as input and that return a

statistiques[205] peuvent être envisagées comme une réponse au problème d'induction dont l'histoire nous rappelle avoir été posé dès le 18e siècle par le philosophe David Hume[206]. Elles visent justement à aborder ce défi de passer du particulier au général de façon rigoureuse en offrant divers procédures et outils[207].

3.2.2 L'approche statistique classique versus l'apprentissage automatique

L'apprentissage automatique se distingue de l'approche « classique » des statistiques[208]. Dans l'approche classique, largement répandue dans la construction des connaissances en sciences sociales[209], un modèle est envisagé *a priori* comme

---

verdict, an evaluation, or some other piece of advice that pertains to the future, or to general states of affairs ».

[205] Christos Mousmoulas, « Probability vs Statistics », (16 décembre 2019), en ligne: *Medium* <https://towardsdatascience.com/probability-vs-statistics-95f221cc74f7>. La statistique est la science de la collecte, l'analyse et l'interprétation des données. Les probabilités concernent l'incertitude. Les statistiques permettent de tirer des conclusions ou prédictions à partir des probabilités. Romeijn, *supra* note 205 : « A method is called statistical, and thus the subject of study in statistics, if it relates facts and hypotheses of a particular kind: the empirical facts must be codified and structured into data sets, and the hypotheses must be formulated in terms of probability distributions over possible data sets ».

[206] Leah Henderson, « The Problem of Induction » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2020 éd, Metaphysics Research Lab, Stanford University, 2020. Ce problème a été mis de l'avant par le philosophe David Hume au 18e siècle et fait l'objet de nombreux débats épistémologiques.

[207] Romeijn, *supra* note 205, à la section 1 : « Much of the philosophy of statistics is about coping with this challenge, by providing a foundation of the procedures that statistics offers, or else by reinterpreting what statistics delivers so as to evade the challenge ».

[208] *Ibid* à la section 5.2.2 : « An entirely different approach to statistics is presented by formal learning theory. This is again a vast area of research, primarily located in computer science and artificial intelligence ». Voir également Danilo Bzdok, Naomi Altman et Martin Krzywinski, « Statistics versus machine learning » (2018) 15:4 Nature Methods 233.

[209] Danilo Bzdok, « Classical Statistics and Statistical Learning in Imaging Neuroscience » (2017) 11 Frontiers in Neuroscience 1. Voir en particulier la Figure 1.

hypothèse de départ et sera testé sur un échantillon dans le but de tirer des conclusions générales ou confirmer l'effet d'une variable sur une autre[210]. En apprentissage automatique, le modèle est dérivé directement des données, sans *a priori*[211], ceci par l'identification des structures (*patterns*) dans les données[212]. C'est donc une approche heuristique, qui utilise la « force brute » des données[213] pour trouver la meilleure approximation d'un phénomène[214]. Les distinctions entre ces deux approches font peu l'objet de publications scientifiques[215] et bien que ni l'une ni l'autre ne soit supérieure

---

[210] Bzdok, Altman et Krzywinski, *supra* note 209 : « Statistical methods have a long-standing focus on inference, which is achieved through the creation and fitting of a project-specific probability model. The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a "true" effect that is unlikely to result from noise ».

[211] Bzdok, *supra* note 210 : « Tools from ClSt [classical statistics] therefore typically assume that the data behave according to certain known mechanisms, whereas StLe [statistical learning] exploits algorithmic techniques to avoid many a-priori specifications of data-generating mechanisms ».

[212] Romeijn, *supra* note 205 : « The discipline is here mentioned briefly, as another example of an approach to statistics that avoids the choice of a statistical model altogether and merely identifies patterns in the data ».

[213] Bzdok, *supra* note 210 fig 1 : « StLe [statistical learning] takes a brute-force approach to model the output of the black box [...] from its input [...] while making a possible minimum of assumptions ».

[214] Bzdok, *supra* note 210 : « ClSt [classical statistics] tends to be more analytical by imposing mathematical rigor on the phenomenon, whereas StLe [statistical learning] tends to be more heuristic by finding useful approximations ».

[215] *Ibid* : « There is currently a scarcity of scientific papers and books that would provide an explicit account on how concepts and tools from classical statistics and statistical learning are exactly related to each other ».

ou universellement applicable[216], chacune peut donner lieu à des conclusions différentes[217].

La nouvelle approche des statistiques mise en œuvre dans les algorithmes d'apprentissage automatique est formalisée au sein d'une branche de l'informatique appelée théorie de l'apprentissage (*computational learning theory* ou *formal learning theory*)[218]. La théorie de l'apprentissage vise à formaliser de façon théorique divers aspects des algorithmes d'apprentissage[219], par exemple les types de problèmes qu'ils peuvent résoudre, leur pouvoir prédictif, l'efficacité avec laquelle ils utilisent le volume et le type des données d'entrées (*input*)[220], le nombre d'exemples requis par un algorithme afin de résoudre un problème donné[221], leur robustesse aux erreurs dans les

---

[216] *Ibid* : « Neither ClSt [classical statistics] or StLe [statistical learning] nor any of the other categories of statistical models can be considered generally superior. This relativism is captured by the so-called no free lunch theorem4 (Wolpert, 1996): no single statistical strategy can consistently do better in all circumstances (cf. Gigerenzer, 2004) ».

[217] *Ibid*. L'auteur relate plusieurs études de cas en neuroscience où l'approche statistique infuence le résultat. Il conclut en paraphrasant Feyerabend : « The goal and permissible conclusions of a neuroscientific investigation are therefore conditioned by the adopted statistical framework (cf. Feyerabend, 1975) ».

[218] Oliver Schulte, « Formal Learning Theory » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2018 éd, Metaphysics Research Lab, Stanford University, 2018.

[219] Sally A Goldman, « Computational learning theory », (1991), en ligne: <https://dl.acm.org/doi/pdf/10.5555/1882757.1882783>.

[220] *Ibid* aux pp 26-2; Jordan et Mitchell, « Machine learning », *supra* note 158. Certains problèmes dans certaines conditions sont insolubles : « Given a learning problem with a given volume of training data, is it possible to design a successful algorithm or is this learning problem fundamentally intractable? ».

[221] Goldman, *supra* note 220.

données d'entrées, etc[222]. La théorie de l'apprentissage permet, entre autres, de comparer les performances de divers algorithmes autrement que par la voie empirique qui, elle, prête à interprétation[223].

L'apprentissage automatique se situe donc à l'intersection de l'informatique, des statistiques et de la théorie de l'apprentissage.

3.2.3 Les enjeux liés à l'utilisation des statistiques

3.2.3.1 Biais cognitifs et difficultés sémantiques

Les statistiques sont notoires pour poser des défis de compréhension dus à nos biais cognitifs[224], risquant ainsi de mener à une surestimation ou une sous-estimation du phénomène sous analyse[225]. Les travaux des psychologues Kahneman et Tversky démontrent nos biais dans l'évaluation des probabilités, par exemple notre aversion naturelle au risque et notre propension à déformer les probabilités dans le cadre de

---

[222] Jordan et Mitchell, « Machine learning », *supra* note 158 : « How robust is the algorithm to errors in its modeling assumptions or to errors in the training data? ».

[223] Goldman, *supra* note 220 aux pp 26-2. À propos de comparaisons empiriques : « it is hard using such evaluations to make meaningful comparisons among competing learning algorithms ».

[224] Steven Pinker, « GENED 1066: Rationality », (2019), en ligne: *Harvard* <https://harvard.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx#folderID=%2255a37adc-eaae-4aa6-8a06-ab25015a4ee8%22&page=0> à la session « Statistical Decision Making ». En particulier, les travaux de Kahneman et Tversky nous éclairent sur les biais qui teintent notre jugement de phénomènes incertains. Daniel Kahneman et Amos Tversky, « On the psychology of prediction » (1973) 80:4 Psychological Review 237.

[225] Andrea Roth, « The Use of Algorithms in Criminal Adjudication » dans *The Cambridge Handbook of the Law of Algorithms* Cambridge Law Handbooks, Cambridge University Press, 2020, 407 à la p 427.

décisions risquées[226]. Notre difficulté à interpréter les statistiques peut donner lieu à de nombreuses erreurs[227]. Ainsi, plusieurs cas d'utilisation erronée des statistiques dans le cadre d'affaires légales ont été documentés [228] , certains donnant lieu à des renversements de verdict[229]. Non seulement les statistiques posent-elles des enjeux d'interprétation mais finalement comptent assez peu dans notre appréciation d'un phénomène, selon Tim Miller[230]. En effet, au-delà des chiffres ce sont surtout les relations causales ou associatives qui nous permet de comprendre un phénomène[231].

La sémantique des statistiques (par exemple, la distinction entre les termes « probable » et « vraisemblable » n'a pas la même envergure en terme statistique qu'en français courant), particulièrement dans un pays bilingue où les traductions sont constantes, porte à confusion et peut créer des effets indésirables, tel que l'illustrent Crispino et ses coauteurs dans deux arrêts nord-américains[232]. Ainsi, divers juristes

---

[226] Frédéric Martinez, « L'individu face au risque : l'apport de Kahneman et Tversky » (2010) N° 161:3 Idees economiques et sociales 15.

[227] Norman Fenton et Martin Neil, « Avoiding Probabilistic Reasoning Fallacies in Legal Practice using Bayesian Networks » (2011) 36 Australian Journal of Legal Philosophy. Les auteurs décrivent des erreurs courantes d'interprétation des statistiques à la section 2 ainsi que leur répercussion en cour à la section 3.

[228] *Ibid*.

[229] Un exemple est le cas de Sally Clarke au Royaume-Uni où l'inculpée fut trouvé coupable à tort pour le meurtre de ses enfants en 1999 suite à un calcul erroné de probabilités. Le verdict a été par la suite renversé en 2003. *Ibid*; Norman Fenton, « Improve statistics in court » (2011) 479:7371 Nature 36.

[230] Tim Miller, « Explanation in Artificial Intelligence: Insights from the Social Sciences » (2018) arXiv:170607269 [cs], en ligne: <http://arxiv.org/abs/1706.07269>, arXiv: 1706.07269.

[231] *Ibid*.

[232] Frank Crispino, Cyril Muehlethaler et Liv Cadola, « Vraisemblable n'est pas probable. De la nécessité d'une sémantique rigoureuse entre scientifiques et juristes » (2020) 2:1 Revue canadienne de

prônent un encadrement de l'approche statistique utilisée en cour ou préconisent un standard terminologique entre les experts en sciences criminelles et les juristes[233]. De ce fait, un groupe constitué de juristes et statisticiens de la UK Royal Statistical Society a produit un guide visant les juges, avocats et experts en sciences criminelles pour mieux appréhender les statistiques au sein des tribunaux[234]. La Figure 3 illustre l'équivalent verbal d'un ratio entre deux probabilités utilisées à la Cour en Nouvelle-Zélande[235].

Likelihood Ratio	Verbal Equivalent
1	Inconclusive
1 to 10	Slightly support the prosecution proposition
10 to 100	Moderately supports
100 to 1,000	Strongly supports
1000 to 1,000,000	Very strongly supports
Greater than 1,000,000	Extremely strongly supports

Figure 3 - Équivalence verbale du rapport de vraisemblance[236]

justice et droit 95. Les auteurs réfèrent aux arrêts « US v Gissantaner » et « R v Joel France », aux notes 6 et 7.

[233] Fenton et Neil, *supra* note 228; Crispino, Muehlethaler et Cadola, *supra* note 233. Les auteurs mentionnent qu'une terminologie commune est préconisée ou standardisée en Europe et en Australie. Voir la note 16.

[234] Fenton et Neil, *supra* note 228 Voir la note 8 : Aitken, C., Roberts, P. & Jackson, G. Fundamentals of Probability and Statistical Evidence in Criminal Proceedings (Royal Statistical Society, 2010).

[235] Moez Ajili, *Reliability of voice comparison for forensic applications*, thèse de doctorat en informatique, Université d'Avignon et des Pays de Vaucluse, 2017 [non publiée] à la p 41. Ce ratio est appelé « ratio de similitude ». Il est utilisé dans plusieurs applications d'apprentissage automatisé.

[236] L'image est tirée de ce document qui réfère à l'équivalence verbale du ratio de similitude au tribunal en Nouvelle-Zélande : *Ibid* à la p 43.

Le témoignage expert devrait s'accompagner d'un guide terminologique afin de favoriser une même interprétation des probabilités par les acteurs de la cour.

3.2.3.2 L'approche bayésienne

Certains algorithmes d'apprentissage utilisent une approche statistique bayésienne et d'autres une approche fréquentiste[237]. Nous distinguons dans ce qui suit les deux approches dont les résultats peuvent diverger de façon considérable[238]. L'affaire *People vs Puckett*[239] aux États-Unis est représentative de cette divergence. En effet, dans cette affaire les résultats d'une identification par analyse d'ADN selon l'approche fréquentiste ou bayésienne pouvait diverger d'un facteur de 1 million[240].

L'approche fréquentiste implique une complète indépendance d'un événement par rapport au même événement dans le passé[241]. L'exemple d'un lancer de dé est typique de l'approche fréquentiste : chaque lancer de dé a une chance sur 6 de tomber sur le 3, indépendamment des lancers précédents. L'approche bayésienne prend en

---

[237] Pedro Domingo, « The Five Tribes of Machine Learning », (2015), en ligne: *Association for Computing Machinery (ACM)* <https://learning.acm.org/binaries/content/assets/leaning-center/webinar-slides/2015/five-tribes-ml_112415.pdf>.

[238] *Manuel scientifique*, *supra* note 39 à la p 82.

[239] *Ibid*. L'affaire est mentionnée à la note 37.

[240] *Ibid*.

[241] *Ibid* aux pp 80, 148.

considération la probabilité d'une hypothèse *avant* le test (hypothèse *a priori*), laquelle sera mise à jour au fur et à mesure des nouvelles observations[242]. Ce mode de calcul doit être considéré dans certains contextes, tel qu'illustré par l'exemple suivant[243] .

On cherche à comparer deux enregistrements pour savoir si le locuteur inconnu sur un enregistrement est identique au locuteur de l'autre enregistrement qui, lui est connu. Un test de reconnaissance de la voix est effectué et produit comme résultat un *rapport de vraisemblance* de 1/1000. Le rapport de vraisemblance est un rapport entre deux probabilités que l'on exprime comme suit :

$$Rapport\ de\ vraisemblance\ = \frac{P(E|H1)}{P(E|H2)}$$

Deux hypothèses s'affrontent :

- L'hypothèse 1 (H1) est que les voix sur les deux enregistrements sont identiques.
- L'hypothèse 2 (H2) est que les voix sur les deux enregistrements sont distinctes.

---

[242] *Ibid*.

[243] L'exemple et les explications sont tirés de cet ouvrage : Geoffrey Stewart Morrison, Cuiling Zhang et Ewald Enzinger, « Forensic speech science » dans Ian Freckelton et Hugh Selby, dir, *Expert Evidence*, Sydney, Australia, 2019 à la p 19.

Les hypothèses du rapport de vraisemblance correspondent en général aux thèses présentées par les parties opposées en cour[244]. Le rapport de vraisemblance exprime la force de chaque hypothèse où :

- E représente les caractéristiques acoustiques observées de la voix inconnue.
- P(E|H1) est la probabilité d'obtenir E avec l'hypothèse 1 (H1)
- P(E|H2) est la probabilité d'obtenir E avec l'hypothèse 2 (H2)

Un rapport plus grand que 1 exprime que les observations E ont plus de chances de survenir avec deux voix identiques que deux voix distinctes. Un rapport plus petit que 1 exprime que les observations E ont plus de chances de survenir avec deux voix distinctes que deux voix identiques.

Une autre façon de concevoir le rapport de vraisemblance est de considérer le numérateur comme étant le degré de *similitude* et le dénominateur comme étant le degré de *typicité (typicality)*. En effet, la similitude seule n'indique pas grand-chose en soi, deux enregistrements peuvent avoir des caractéristiques acoustiques similaires mais tellement typiques que n'importe quels deux enregistrements pris au hasard seraient tout aussi similaires. Ainsi, il faut mesurer également la *typicité* des caractéristiques au sein d'une *population de référence*[245]. La population de référence doit être pertinente[246]. Dans notre exemple, elle doit avoir les mêmes langue, accent et genre que les voix sur les enregistrements. En général, plus la similarité est forte et la

---

[244] Jean-François Bonastre et al, « Forensic Speaker Recognition: Mirages and Reality » dans *Individual Differences in Speech Production and Perception*, Peter Lang International Academic Publishers, 2015, 255 à la p 262. Ces thèses pourront varier tout au long du procès.

[245] Morrison, Zhang et Enzinger, *supra* note 244 à la p 19.

[246] Bonastre et al, *supra* note 245.

typicité faible, plus l'hypothèse 1 doit être considérée. Inversement, plus la similarité est faible et la typicité forte, plus l'hypothèse 2 devra être considérée[247].

Le théorème de Bayes peut être représenté comme suit :

$$\text{Probabilité } a\ posteriori = \text{rapport de vraisemblance} \times \text{probabilité } a\ priori$$[248]

Supposons qu'une analyse acoustique donne un rapport de vraisemblance de 1/1000 dans le contexte d'un crime commis sur une île où habitent 100 personnes, dont le suspect[249]. Quelle serait la probabilité *a priori* que le suspect ait commis le crime ? Avant l'analyse acoustique, la probabilité que le suspect ait commis le crime est de 1/100, si l'on suppose que ce dernier a autant de chance de commettre le crime que toute autre personne sur l'île. Les 99 autres personnes sur l'île ont également une probabilité de 1/100 d'avoir commis le crime. La somme des probabilités que l'une de ces 99 personnes ait commis le crime est de 99 * 1/100 = 99/100. La probabilité *a priori* est donc la probabilité que le suspect soit le meurtrier divisé par la probabilité de cette somme, soit 1/100 divisé par 99/100, ce qui donne 1/99.

En reprenant la formule ci-haut, la probabilité *a posteriori* sera donc de 1/1000, soit le rapport de vraisemblance, multiplié par 1/99, ce qui donne 1/99000 ou 0,001%.

---

[247] Morrison, Zhang et Enzinger, *supra* note 244.

[248] Ken Doyle, *Bayes Theorem. Can Statistics Help Guide a Verdict in the Courtroom?*, The ISHI Report, News from the World of DNA Forensics, International Symposium on Human Identification, 2021.

[249] L'exemple est tiré de Morrison, Zhang et Enzinger, *supra* note 244 à la p 23.

L'hypothèse 2 (les voix provenant des deux enregistrements sont distinctes) doit être favorisée par rapport à l'hypothèse 1 (les voix sont similaires).

Selon Crispino et ses coauteurs, il est inapproprié pour l'expert de soumettre une probabilité *a priori*[250]. Ce dernier ne doit soumettre que le rapport de vraisemblance. C'est le rôle du juge des faits d'estimer la probabilité *a priori* à la lumière des faits[251]. La Commission du droit de l'Ontario, dans une étude au sujet de l'analyse d'ADN dans laquelle l'approche bayésienne est utilisée relève que ce calcul contreviendrait à la présomption d'innocence puisque la probabilité *a priori* ne peut être zéro car dans ce cas la probabilité *a posteriori* serait toujours zéro étant donné qu'un nombre multiplié par zéro donne zéro[252].

Lors de l'évaluation des probabilités, le témoignage d'expert doit rendre explicites les hypothèses considérées, la probabilité *a priori* et la population de référence de l'hypothèse 2.

3.2.3.3 La population de référence

Nous avons vu le rôle de la population de référence dans l'approche bayésienne mais peu importe l'approche, l'interprétation d'une probabilité doit toujours tenir

---

[250] Crispino, Muehlethaler et Cadola, *supra* note 233.

[251] *Ibid*.

[252] Presser et Robertson, *supra* note 115.

compte de la *population de référence*[253]. En effet, une probabilité s'applique nécessairement à un groupe donné[254]. Il existe une myriade de populations de référence potentielles, par exemple selon l'âge, le lieu de résidence, etc. Or, comment définir la population de référence qui convient le mieux dans un contexte donné, sachant que deux populations distinctes pourraient engendrer des résultats distincts[255]? Plus la population de référence est large, plus les résultats pourraient s'avérer incorrects et correspondre à de fausses corrélations[256]. Par exemple, une corrélation entre des difficultés respiratoires et le fait d'habiter Calgary plutôt qu'Edmonton pourrait masquer le fait que c'est plutôt le fait d'habiter au centre-ville versus la banlieue qui est en cause[257]. Dans ce cas, la population de référence doit être raffinée pour distinguer la banlieue du centre-ville.

La population de référence devrait être pertinente selon le cas d'espèce et devrait être explicite dans le témoignage d'expert.

---

[253] On parlera aussi de « classe de référence », voir Russell Brown, « The Possibility of "Inference Causation": Inferring Cause-in-Fact and the Nature of Legal Fact-Finding » (2010) 55:1 RD McGill 1.

[254] *Ibid* à la p 25.

[255] Brown, « The Possibility of "Inference Causation" », *supra* note 254.

[256] *Ibid*.

[257] L'exemple est tiré de Brown, *ibid* à la p 27.

3.2.3.4 L'individualisation

Le droit et la science cherchent tous deux à faire jaillir la vérité[258]. Cependant, tandis que la science cherche habituellement à généraliser les cas particuliers en lois générales, le droit cherche à établir la vérité sur un cas particulier[259]. Ainsi, les statistiques quoiqu'elles puissent servir de socle scientifique à la preuve juridique, permettent de généraliser des phénomènes mais offrent peu de perspective sur un cas d'espèce[260]. Par exemple, assigner une probabilité au fait qu'un homme choisi au hasard gagne un salaire plus élevé que ses consœurs uniquement à partir de statistiques génériques démontrant que le salaire des hommes est plus élevé que celui des femmes manquerait d'exactitude. Certaines techniques statistiques permettent d'affiner les probabilités génériques en fonction de caractéristiques personnelles spécifiques[261].

---

[258] *Barendregt c Grebliunas*, [2022] CSC 22 au para 45; *R c Nikolovski*, [1996] 3 RCS 1197 à la p 1206.

[259] La vérité scientifique et la vérité juridique se distinguent à plusieurs égards, dont notamment quant à la temporalité dans laquelle elles s'inscrivent respectivement. En effet, selon le juge Binnie cité dans l'affaire *Imperial Tobacco*, une affaire en Cour se doit d'être résolue en temps opportun tandis que la science n'a pas cette contrainte. Elle évolue et n'a de cesse de raffiner sa vérité: « The court is a dispute resolution forum, not a free-wheeling scientific inquiry, and the judge must reach a timely decision based on the available information. ». *Imperial Tobacco Canada ltée c Conseil québécois sur le tabac et la santé*, [2019] QCCA 358 au para 746, citant The Honourable Mr. Justice Ian Binnie, « Science in the Courtroom: The Mouse that Roared », (2007) 56 UNBLJ 307 à la p 312.

[260] Brown, « The Possibility of "Inference Causation" », *supra* note 254 à la p 18.

[261] Richard Goldberg, « Epidemiological Uncertainty, Causation, and Drug Product Liability » (2014) 59:4 RD McGill 777. L'auteur aborde la technique de régression logistique pour tenir compte de facteurs de risques individuels dans un contexte épidémiologique.

Dans notre exemple, on pourra considérer le nombre d'années d'expérience, diplôme, poste et domaine de travail de l'homme en question pour personnaliser les probabilités.

> Le témoignage d'expert devrait spécifier les techniques qui tiennent compte des caractéristiques individuelles dans le modèle statistique. Le juge des faits pourra en tenir compte dans l'interprétation de statistiques appliquées à un cas spécifique.

3.3 Le modèle du système

On se rappellera que le modèle du système d'apprentissage automatique est produit à la suite des phases d'entraînement et de tests. Il est le résultat d'un processus d'induction par lequel une hypothèse *a posteriori* est validée empiriquement. Le modèle est un ensemble de variables et de poids, déterminés statistiquement par corrélations entre les données en entrée et le résultat attendu.

3.3.1 Les enjeux du modèle

3.3.1.1 La subjectivité des objectifs

La production d'un modèle implique des choix humains. Les visions et présuppositions des concepteurs et programmeurs entrent en ligne de compte tout au long de la conception et du développement du modèle[262]. En l'occurrence, les objectifs d'apprentissage doivent être définis. Ceux-ci pourraient être basés sur des décisions

---

[262] Leese, Kaufmann et Egbert, *supra* note 17; Jon Kleinberg et al, « Discrimination in the Age of Algorithms » (2018) 10 Journal of Legal Analysis 113.

subjectives passées[263]. Par exemple, Virginia Eubanks relève le cas d'un outil de prédiction de risque de maltraitance juvénile entraîné à partir de dénonciations de maltraitance faites par la communauté, une donnée potentiellement subjective. Ainsi l'objectif de l'outil correspond au risque de dénonciations et non au risque de maltraitance, ce qui jette une lumière bien différente sur l'interprétation des résultats de l'outil[264].

3.3.1.2 La vision subjective du phénomène considéré

Le modèle peut être entièrement conditionné par la vision subjective d'un phénomène. Par exemple, Kaufman illustre les diverses visions politiques du crime sur lesquelles sont fondés les outils de prédictions de crimes[265]. Certaines visions mettent l'emphase sur l'environnement où le crime pourrait avoir lieu, d'autres sur les comportements criminels, et enfin d'autres sur le réseau des relations criminelles[266]. Chaque vision est ainsi encodée dans l'algorithme. Le modèle résultant reflétera un certain point de vue sur la criminalité, point de vue qui influencera les prédictions du système.

---

[263] Jessica Fjeld et al, *Principled Artificial Intelligence : Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, 2020 à la p 47 : « The outcome of interest may be influenced by earlier decisions that are themselves biased ». Les auteurs traitent des risques de perpétuer la discrimination à travers les algorithmes.

[264] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, USA, St. Martin's Press, Inc., 2018 aux pp 143-144. L'auteure donne cet exemple pour mettre en lumière le potentiel discriminatoire des objectifs. Dans ce cas précis, les dénonciations de maltraitance sont subjectives et peuvent contribuer à stigmatiser certains groupes sociaux lorsqu'ils font les frais de dénonciations répétitives et non fondées.

[265] Leese, Kaufmann et Egbert, *supra* note 17.

[266] *Ibid*.

Le témoignage d'expert devrait mettre en lumière les objectifs du modèle et la vision du phénomène encodé dans le modèle.

3.4 Résumé

Les statistiques, fondement de la technique d'apprentissage automatique, posent des défis sémantiques et cognitifs qui peuvent fausser l'interprétation du résultat d'un système. Pour correctement interpréter ce résultat, les hypothèses, la population de référence et les caractéristiques individuelles du cas d'espèce utilisé dans le calcul statistique doivent être considérées. Les objectifs du système d'apprentissage automatique ainsi que le cadre théorique (la vision) qui permet d'atteindre ces objectifs doivent également être considérés.

Après avoir couvert les fondements d'un système d'apprentissage automatique, nous nous penchons au cours du prochain chapitre sur le processus par lequel le système produit un résultat.

CHAPITRE IV

LA FIABILITÉ DU PROCESSUS

4.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de son processus

Ce chapitre concerne la fiabilité du processus qui permet de générer le résultat présenté en preuve. Les questions suivantes pourront contribuer à évaluer ce processus :

- Quelles sont les données et sources de données sur lesquelles le système a été entraîné ainsi que les données qui ont permis d'obtenir le résultat?
- Quelles sont les limites et contraintes du système et comment s'appliquent-elles au cas d'espèce?
- Où intervient l'humain dans le processus de production du résultat et sa formation est-elle adéquate?

4.2 Les enjeux

4.2.1 Les données

Nous avons vu que les données sont au cœur des systèmes d'apprentissage automatique, tant lors de l'entraînement du système, de la validation et des tests du

modèle. Or, les données ne sont pas neutres, elles relèvent de choix humains[267]. Une donnée à l'état brut n'existe pas. Elle est nécessairement transformée pour remplir un objectif et potentiellement modifiée pour se conformer à certaines contraintes, techniques par exemple, pour être traitée par un algorithme[268]. Le choix des données reflète donc des croyances et des valeurs[269].

Le choix des données aura une influence majeure sur le modèle résultant[270], rendant le système vulnérable aux conditions suivantes[271] :

- Les données réelles doivent être **stables** car des variations sont propices aux erreurs[272]. Par exemple, les algorithmes de reconnaissance d'images peuvent être facilement bernés par des variations imperceptibles à l'œil nu[273]. Un

---

[267] Itiel Dror, « The Ambition to be Scientific : Human Expert Performance and Objectivity » (2013) 53:2 Science & Justice: Journal of the Forensic Science Society 81-82 : « Sampling and determining what qualifies as "data" to be used as input to the instrumentation and statistical models are highly influenced by motivational and expectation biases ».

[268] Polyzotis et al, *supra* note 174.

[269] Simon Lindgren, « Hacking Social Science for the Age of Datafication » (2019) 1:1 Journal of Digital Social Research 1 à la p 2. Les données sont toujours manipulées : « The data that we will have at hand are always configured via beliefs, values, and choices that "cook" the data from the very beginning so that they are never in a "raw" state. So, there is no such thing as raw data ».

[270] Leese, Kaufmann et Egbert, *supra* note 17. L'auteur donne plusieurs exemples de biais dans le domaine de la police prédictive.

[271] Les conditions énumérées sont tirées de la littérature sur les algorithmes d'apprentissage profond. Voir Marcus, « Deep Learning », *supra* note 162.

[272] *Ibid* à la section 3.8.

[273] Anh Nguyen, Jason Yosinski et Jeff Clune, « Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images » (2015) arXiv:14121897 [cs], en ligne:

contexte où les règles sont fixes (les jeux en sont de bons exemples) est plus propice au bon fonctionnement de l'algorithme qu'un environnement où les règles changent constamment[274].

- Les données d'apprentissage doivent être **variées**; elles doivent représenter le plus de scénarios possibles[275]. Par exemple, en reconnaissance d'images, un objet présenté selon un point de vue légèrement différent pourrait confondre le système[276].
- Les données doivent être **complètes**. Le problème à résoudre doit être autosuffisant (*self-contained*); on ne peut pas présumer que le système ait de connaissances *a priori* ou des connaissances qu'il ne peut apprendre à partir des données disponibles[277]. Un exemple de connaissances *a priori* serait les lois de la physique dans une application de prévisions météorologiques. Certaines connaissances *a priori* qui relèvent pour les humains du « gros bon sens » sont difficiles à intégrer dans les systèmes d'apprentissage. Ce genre de problème en est un exemple : « qui est plus grand, le prince William ou son bébé le Prince George? »[278].

<http://arxiv.org/abs/1412.1897>, arXiv: 1412.1897. Les auteurs mentionnent plusieurs exemples où la reconnaissance d'images est erronée.

[274] Marcus, « Deep Learning », *supra* note 162 à la section 3.8. L'auteur donne un exemple dans le domaine épidémiologique où les performances de l'algorithme Google Flu Trends ont radicalement diminué d'une année à l'autre.

[275] Dickson, *supra* note 170.

[276] *Ibid*.

[277] Marcus, « Deep Learning », *supra* note 162 à la section 3.6.

[278] *Ibid* à la section 3.6.

Le manque de données de qualité est l'un des principaux obstacles au déploiement de modèles performants, selon un sondage de Rackspace Technologie conduit en 2021 auprès de plus de 1800 experts de l'industrie[279]. L'accès aux données de qualité constitue donc un facteur distinctif et un avantage commercial majeur.

> Le témoignage d'expert devrait fournir les sources des données d'apprentissage et des données réelles pour s'assurer qu'elles sont stables, complètes et variées.
>
> Il devrait identifier les limites et contraintes du système et si celles-ci le rendent vulnérable dans le cas d'espèce.

4.2.2 L'intervention humaine

Un système peut être perçu comme un binôme humain-machine et non une machine seule[280]. Dans cette optique, l'humain interagit avec la machine à différentes étapes incluant la prise de décision finale dont il est l'ultime responsable, selon Ezer et

---

[279] Le sondage est disponible à partir de ce site : Rackspace Technology, « AI and Machine Learning Research Report », (26 janvier 2021), en ligne: <https://www.rackspace.com/solve/succeeding-ai-ml>; James Vincent, « The Biggest Headache in Machine Learning? Cleaning Dirty Data Off the Spreadsheets », (1 novembre 2017), en ligne: *The Verge* <https://www.theverge.com/2017/11/1/16589246/machine-learning-data-science-dirty-data-kaggle-survey-2017>. L'article réfère à un sondage de la compagnie Kaggle (qui offre des données à diverses industries). Cette citation du fondateur de Kaggle illustre avec humour l'effort requis pour obtenir des données de qualité : « There's the joke that 80 percent of data science is cleaning the data and 20 percent is complaining about cleaning the data » .

[280] Neta Ezer et al, « Trust Engineering for Human-AI Teams » (2019) 63 Proceedings of the Human Factors and Ergonomics Society Annual Meeting 322. Un axe de recherche sur l'interaction entre l'humain et la machine est appelé « trust engineering ». La recherche se penche entre autres sur la formation (ou apprentissage) tant de la machine que de l'humain.

ses coauteurs[281]. À ce titre, l'humain devrait connaître les limites du système et les conditions propices à l'erreur[282] ou ce à quoi Hoffman réfère comme étant le « modèle mental » avec lequel l'humain se représente le système[283].

C'est par le biais d'une formation et diverses interactions avec le système que l'humain développera un modèle mental fidèle au fonctionnement du système, ce qui lui permettra d'évaluer adéquatement les recommandations du système (et augmenter les performances du binôme)[284]. La formation pourra l'aider à contrer le *biais d'automatisation*[285] qui réfère à la propension naturelle de l'humain à accepter sans

---

[281] Pour plus d'information sur la notion du binôme humain-machine, voir cet article qui élabore sur le « trust engineering » : *Ibid*.

[282] Gagan Bansal et al, « Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance » (2019) 7:1 Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 2. Les auteurs réfèrent aux limites du système ainsi : « error boundaries ».

[283] Robert R Hoffman et al, « Metrics for Explainable AI: Challenges and Prospects » (2019) arXiv:181204608 [cs], en ligne: <http://arxiv.org/abs/1812.04608>, arXiv: 1812.04608.

[284] *Ibid* fig 1. Meilleur est le modèle mental, meilleure sera la performance du binôme humain-machine.

[285] Ashley Deeks, « The Judicial Demand for Explainable Artificial Intelligence, Virginia Public Law and Legal Theory Research Paper 2019-51 » (2019) SSRN, en ligne: <https://papers.ssrn.com/abstract=3440723>.

vérifier les résultats produits par une machine[286]. Ainsi, l'humain sera plus enclin à outrepasser les recommandations de l'algorithme.

Le témoignage expert devrait distinguer l'intervention humaine de l'automatisation lors du processus de production du résultat et confirmer que le niveau d'expertise de l'humain est adéquat.

4.3 Résumé

Les sources premières d'un système d'apprentissage automatique sont les données. Les données utilisées dans la conception du système doivent être soigneusement choisies afin de refléter au mieux la réalité. Un système informatique sera optimal à l'intérieur de certaines contraintes et limites qui devraient être maîtrisées par l'utilisateur du système.

Après nous être penchés sur les fondements du système d'apprentissage automatique et le processus, automatisé ou manuel, de production du résultat, nous nous penchons sur l'explication de ce résultat.

---

[286] *Ibid* à la p 1846. L'auteur nomme cette tendance « automation bias ».

CHAPITRE V

L'EXPLICATION DU RÉSULTAT

5.1 Les questions que les juristes doivent se poser devant un système en lien avec l'explication du résultat obtenu

Les questions portent sur la façon dont le résultat spécifique a été produit par le système[287] :

- Comment explique-t-on le résultat spécifique pour ce cas d'espèce?
- Quelle est l'incertitude liée au résultat?
- Quels sont les données, leurs valeurs et poids qui ont permis de produire le résultat?
- Ces données sont-elles exactes et cohérentes avec l'objectif du système?
- Le résultat aurait-il été différent si l'on modifiait une donnée x?
- Quel est le plus petit changement susceptible de modifier un résultat favorable en résultat défavorable (ou vice versa)?

---

[287] Plusieurs des questions suivantes sont tirées de ce rapport : Finale Doshi-Velez et al, « Accountability of AI Under the Law: The Role of Explanation » (2019) arXiv:171101134 [cs, stat], en ligne: <http://arxiv.org/abs/1711.01134> aux pp 2-3, arXiv: 1711.01134.

5.2 L'explication algorithmique

Bien que la transparence soit une notion complexe et polysémique qui reçoit des définitions diverses dans des contextes variés[288], dans le contexte d'un algorithme elle renvoie à notre capacité à comprendre les décisions prises par l'algorithme et à nous assurer qu'il fasse bien ce qu'il est censé faire[289]. L'aspect épistémologique fondamental de la transparence est l'intelligibilité[290] définie, toujours dans le contexte d'un algorithme, comme étant la possibilité « de comprendre son fonctionnement et de vérifier s'il satisfait bien les propriétés désirées »[291]. L'intelligibilité est donc « une forme d'explicabilité fondamentale »[292].

La transparence et son contraire, l'opacité, sont donc liées au concept d'explication dont l'objectif est la compréhension du traitement algorithmique[293]. L'explication permet de traduire les concepts techniques et les résultats des algorithmes en termes compréhensibles et dans un format propice à l'évaluation[294]. En d'autres

---

[288] Maël Pégny et Issam Ibnouhsein, « Quelle transparence pour les algorithmes d'apprentissage machine ? » (2018) 32 Revue d'intelligence artificielle 447 à la p 4.

[289] *Ibid* à la p 3.

[290] *Ibid*.

[291] *Ibid* à la p 8.

[292] *Ibid*.

[293] *Ibid* à la p 4.

[294] Fjeld et al, *supra* note 264 à la p 42.

mots l'explication est la description du processus par lequel les données en entrées ont engendré les résultats obtenus en sortie[295].

5.3 Les fonctions de l'explication : de la fiabilité du système au processus judiciaire

L'explication est une condition essentielle pour déboguer ou réaliser un audit sur un système[296]. L'explication est étroitement liée à la détection d'erreurs, à la détection de biais[297] et, de façon plus générale, à l'évaluation des résultats produits par le modèle[298] et à la fiabilité du système[299].

La communauté juridique reconnaît largement le rôle de l'explication algorithmique dans le processus judiciaire, particulièrement dans un contexte pénal où

---

[295] Doshi-Velez et al, « Accountability of AI Under the Law », *supra* note 288 aux pp 2-3.

[296] Christoph Molnar, « Interpretable Machine Learning: A Guide for Making Black Box Models Explainable », (2021), en ligne: <https://christophm.github.io/interpretable-ml-book/index.html> ch 2.1.

[297] CE, Commission, *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen, et au Comité des régions: Façonner l'avenir numérique de l'Europe,* Bruxelles, 2020 n 32. À propos de la recherche sur l'explication algorithmique : « In order to increase transparency and minimise the risk of bias or error, AI systems should be developed in a manner which allows humans to understand (the basis of) their actions ».

[298] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You?": Explaining the Predictions of Any Classifier » (2016) arXiv:160204938 [cs, stat], en ligne: <http://arxiv.org/abs/1602.04938>, arXiv: 1602.04938. Voir la section 2 « The Case for Explanations ».

[299] Fjeld et al, *supra* note 264 aux pp 37, 42. Le concept d'explication recoupe celui de fiabilité (reliability), défini par les auteurs comme suit : « a system that is reliable is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorized third parties. »; Molnar, *supra* note 297 ch 2.1. L'explication favorise les caractéristiques suivantes : « [r]eliability or [r]obustness : [e]nsuring that small changes in the input do not lead to large changes in the prediction ».

les enjeux sont majeurs[300]. Le professeur Cofone, auteur d'une importante proposition de réforme de la LPRPDE[301], décrit au paragraphe 4b certaines implications du droit à l'explication :

> [Le droit à l'explication] favorise les droits en matière de responsabilité en aidant les individus à comprendre les décisions qui les concernent. Elle protège d'autres droits énoncés dans la LPRPDE (y compris des droits recommandés dans le présent rapport), comme *le droit à la contestation, au consentement éclairé ainsi que le droit de déceler et de corriger des renseignements personnels erronés*, dont les inférences. Le droit à une explication protège également les droits de la personne, comme *le droit de ne pas être soumis à des décisions discriminatoires* [nos italiques][302].

L'explication revêt donc de multiples rôles. Elle permet de justifier et légitimer une décision[303] et en ce sens elle contribue à l'intégrité de la cour[304]. Elle contribue à l'équité du processus judiciaire[305] en permettant entre autres d'accéder aux données utilisées pour parvenir à une décision et de contester celle-ci[306]. Elle contribue à la

---

[300] Doshi-Velez et al, « Accountability of AI Under the Law », *supra* note 288 à la p 6.

[301] Ignacio Cofone, *Propositions stratégiques aux fins de la réforme de la LPRPDE élaborées en réponse au rapport sur l'intelligence artificielle*, Commissariat à la protection de la vie privée du Canada, 2020; *La Loi sur la protection des renseignements personnels et les documents électroniques*, LC 2000, c 5.

[302] Cofone, *supra* note 302 au para 4b.

[303] Selbst et Barocas, *supra* note 36.

[304] Deeks, *supra* note 286 à la p 1845.

[305] *Ibid* à la p 1846.

[306] Selbst et Barocas, *supra* note 36; Daniel Kluttz, Kohli, Nitin, et Mulligan, Deirdre K, « Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the

protection de la vie privée[307] et à prévenir la discrimination[308]. De façon plus générale, elle favorise l'autodétermination en nous donnant la possibilité d'agir sur une décision par la compréhension des paramètres qui la compose[309]. Finalement, elle aide à contrer la tendance de l'humain à accepter sans vérifier les résultats produits par une machine, c'est-à-dire le biais d'automatisation que nous avons évoqué ultérieurement[310].

5.4 Les enjeux de l'explication

5.4.1 Complexité des modèles

L'explication des résultats d'un système par apprentissage automatique pose des défis particuliers dus à la complexité potentielle de leurs modèles, en particulier du nombre élevé de variables en jeu[311]. Les modèles d'apprentissage profond peuvent faire

---

Professions » dans *After the Digital Tornado* Cambridge University Press, Kevin Werbach, 2020, 137 : « Contestability, the ability to contest decisions [...] is one of the interests that transparency serves ».

[307] Valérie Beaudouin et al, « Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach » (2020) arXiv:200307703 [cs], en ligne: <http://arxiv.org/abs/2003.07703> à la p 27, arXiv: 2003.07703.

[308] *Ibid* à la p 29. Les auteurs réfèrent à l'affaire NJCM v. the Netherlands : « The court analyzed the algorithm under the EU test of proportionality, and found that the lack of explanation for the computer-generated risk reports prevented individuals from being in a position to challenge the reports, and prevented the court from verifying the absence of discrimination ».

[309] Selbst et Barocas, *supra* note 36.

[310] Deeks, *supra* note 286 à la p 1846.

[311] Selbst et Barocas, *supra* note 36 à la p 1094. Les auteurs distinguent 4 propriétés mathématiques de la complexité sur lesquelles nous élaborons à la section 6.3 : « Four mathematical properties related to model complexity are linearity, monotonicity, continuity, and dimensionality. The dimensionality of a model is the number of variables it considers ».

appel à des millions de variables[312], surpassant largement les capacités cognitives humaines[313]. Ainsi, on qualifiera les modèles complexes d'opaques et on parlera volontiers de leur manque de transparence et des défis que pose leur explicabilité. Nous verrons dans le chapitre suivant les compromis requis entre complexité et explicabilité.

5.4.2 Les fausses corrélations

Nous avons vu que les algorithmes d'apprentissage détectaient des liens entre les données. Ces liens sont des corrélations, que l'on doit distinguer des liens de causalité[314]. Une corrélation est la tendance de deux variables à varier ensemble[315]. Par exemple, plus un homme est grand, plus son poids est élevé; il y a une corrélation *positive* entre la taille et le poids car les deux variables augmentent ensemble. Un exemple de corrélation *négative* est la relation entre l'altitude et la température : plus l'altitude est élevée, plus la température baisse. Le lien de causalité quant à lui implique que lorsqu'un évènement survient (la cause), un autre évènement (l'effet) survient également en succession ou simultanément[316]. Par exemple, gratter une allumette *cause* un feu[317]. Un lien de causalité implique nécessairement une corrélation entre deux

---

[312] Tirthajyoti Sarkar, « Google's New "Explainable AI" (xAI) Service », (2 janvier 2020), en ligne: *Medium* <https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773> : « DL models use millions of parameters and create extremely complex and highly nonlinear internal representations of the images or datasets that are fed to them ».

[313] Selbst et Barocas, *supra* note 36 : « Humans have no way to visualize models with more than three dimensions ».

[314] Eubanks, *supra* note 265 à la p 144.

[315] Steven Pinker, *supra* note 225 à la session 7 « Correlation and Causation ».

[316] The Editors of Encyclopaedia Britannica, « Causation » dans *Encyclopedia Britannica*, 2009.

[317] Steven Pinker, *supra* note 225. L'exemple est tiré de la session 7 « Correlation and Causation » à 00h:00m:37s où Pinker souligne qu'un évènement peut survenir lorsque de multiples conditions sont

événements tandis que la corrélation n'implique pas la causalité[318]. Les algorithmes ne peuvent pas, à l'heure actuelle, détecter des liens de cause à effet[319].

Un exemple classique de la distinction entre corrélation et causalité est celui de la corrélation entre les attaques de requins et la consommation de crème glacée[320]. Malgré la corrélation, il serait mal avisé de bannir les kiosques de crème glacée pour prévenir les attaques de requins! Les deux variables augmentent ensemble mais c'est une troisième variable qui les influence, soit la hausse de température lorsqu'elle devient propice à la baignade et à la consommation de crème glacée. Il peut être difficile, selon le contexte, de distinguer les corrélations de l'effet du hasard car les évènements de pur hasard (sans aucune corrélation) ont une tendance naturelle à se regrouper[321]. Des corrélations peuvent *toujours* être identifiées *post hoc* et s'avérer non-pertinentes au phénomène sous analyse[322]. Ainsi, le risque d'identifier des

---

rencontrées, lesquelles ne sont pas toutes identifiées comme la cause de l'évènement. Dans l'exemple, le feu est conditionnel à la présence d'oxygène, l'absence d'humidité, l'absence de vent, etc.

[318] Paul Weirich, « Causal Decision Theory » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

[319] Ben Dickson, « The Book of Why: Exploring the missing piece of artificial intelligence », (9 décembre 2019), en ligne: *TechTalks* <https://bdtechtalks.com/2019/12/09/judea-pearl-the-book-of-why-ai-causality/>; Will Knight, « If AI's So Smart, Why Can't It Grasp Cause and Effect? » *Wired* (3 septembre 2020), en ligne: <https://www.wired.com/story/ai-smart-cant-grasp-cause-effect/>.

[320] Viginia Eubanks fait référence à cet exemple. Eubanks, *supra* note 265 aux pp 144-145.

[321] Steven Pinker, *supra* note 225 à la session 4 « Probability and Randomness Part 1 » à 01h00m:00s.

[322] *Ibid* à la session 4 « Probability and Randomness Part 1 » à 00h50m:00s. L'auteur donne en exemple les nombreuses corrélations que l'on peut trouver entre les présidents Abe Lincoln et John F Kennedy : les deux présidents ont été élus en '60 (1860 et 1960 respectivement), les deux ont été victime d'un attentat par balle, à la tête, un vendredi, en présence de leur femme, etc. Voir également Matthew Stewart, « The Limitations of Machine Learning », (29 juillet 2019), en ligne: *Towards Data Science* <https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6> : « spurious correlations [...] are usually obtained by p-hacking (looking through mountains of data until a correlation

corrélations là où il n'y a que coïncidences peut être élevé[323], surtout lorsqu'il n'y a pas de mécanisme adéquat et systématique de rejet; le modèle est ajusté jusqu'à ce qu'une similitude soit obtenue[324].

> Dans un modèle complexe, les corrélations ne sont pas toujours accessibles mais lorsque certaines corrélations vont au cœur de ce que la preuve tend à démontrer, le témoignage expert devrait être en mesure d'expliquer ces corrélations entre les données et le résultat afin de confirmer la cohérence du modèle.

5.4.3 Les coûts et le secret commercial

La production d'une explication implique des coûts de production et de mise en place[325]. Ceux-ci incluent des coûts directs, par exemple, les coûts de stockage des

---

showing statistically significant results is found). These are not true correlations and are just responding to the noise in the measurements ».

[323] Pallab Ghosh, « AAAS: Machine learning "causing science crisis" », *BBC News* (16 février 2019), en ligne: <https://www.bbc.com/news/science-environment-47267081> : « machine learning algorithms have been developed specifically to find interesting things in datasets and so when they search through huge amounts of data they will inevitably find a pattern ».

[324] Steven Pinker, *supra* note 225 à la session 4 « Probability vs Randomness ». L'auteur explique le phénomène du « Texas sharp shooter fallacy » où un tireur d'élite texan tire au fusil et par la suite dessine autour du trou de balle une cible. Le modèle de l'apprentissage automatique est confirmé à partir d'un résultat.

[325] Jenna Burrell, « How the Machine "Thinks" : Understanding Opacity in Machine Learning Algorithms » (2016) 3:1 Big Data & Society 2053951715622512. L'article présente les enjeux d'opacité de l'apprentissage automatique. Voir également David Gunning et David Aha, « DARPA's Explainable Artificial Intelligence (XAI) Program » (2019) 40:2 AI Magazine 44. Le programme de recherche DARPA sur l'xAI présente trois facettes de recherche : produire des modèles interprétables, produire des méthodes de visualisation ou représentation d'une explication et comprendre les mécanismes cognitifs à l'oeuvre dans l'interprétation d'une explication.

données impliquées dans une décision[326]. Ils peuvent inclure des coûts indirects, tels que l'amenuisement de la capacité d'innovation technologique et de la compétitivité commerciale[327], en l'occurrence lorsque l'explication révèle certains aspects du secret commercial[328]. Selon le Commissaire à l'information du Royaume-Uni, le coût est le principal frein à la mise en place de systèmes explicables[329]. Finalement, les coûts de l'explication se mesurent également en termes de performance. En effet, l'explication et la performance peuvent parfois être considérées antinomiques : plus le degré d'explicabilité augmente moins les résultats du système seraient précis, ce qui implique des compromis entre ces deux caractéristiques souhaitables[330]. Selon Selbst et Barocas, la recherche en *intelligence artificielle explicable*, sur laquelle nous nous penchons à la prochaine section, est principalement mue par cette tension entre performance et interprétabilité[331].

---

[326] Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 aux pp 35, 38. Notons qu'une stratégie de stockage peut parfois entrer en conflit avec les lois existantes sur la protection des données personnelles lorsqu'il s'agit par exemple de données biométriques.

[327] Doshi-Velez et al, « Accountability of AI Under the Law », *supra* note 288 à la p 2.

[328] *Ibid*.

[329] Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 aux pp 35, 38 : « United Kingdom's Information Commissioner's Office ».

[330] *Ibid* à la p 36.

[331] Selbst et Barocas, *supra* note 36 à la p 1110.

5.5 L'intelligence artificielle explicable

Le domaine de recherche qui traite spécifiquement de l'explication algorithmique en intelligence artificielle est *l'intelligence artificielle explicable* (*Explainable Artificial Intelligence* ou *xAI*). L'intelligence artificielle explicable englobe trois axes de recherche : « la production de modèles plus explicables, la production d'interfaces utilisateurs plus intelligibles et la compréhension des mécanismes cognitifs à l'œuvre pour produire une explication satisfaisante »[332]. L'intelligence artificielle explicable se situe à l'intersection de la philosophie, de la psychologie, des sciences cognitives et de l'informatique[333].

5.5.1 Portée de l'explication

Il existe plusieurs techniques d'explication pour les systèmes d'apprentissage automatique et plusieurs façons de les classer[334]. Le choix d'une technique dépend

---

[332] Villani et al, *supra* note 25 à la p 145. Le rapport Villani recommande ces trois axes de recherche. Il se base sur les trois axes du programme DARPA de la défense américaine. Le programme DARPA sur l'IA explicable regroupe au moins 13 projets distincts, voir Gunning et Aha, *supra* note 326. La défense américaine considère l'explication de l'IA comme un enjeu majeur (« key stumbling block ») alors des milliards sont investis dans les applications militaires qui utilisent des techniques d'apprentissage automatique : Will Knight, « The Dark Secret at the Heart of AI », (2017), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>.

[333] Miller, « Explanation in Artificial Intelligence », *supra* note 231 à la p 4. Selon Miller, l'IA explicable aurait grand avantage à construire sur les acquis des sciences sociales qui depuis des décennies (en psychologie et en sciences cognitives) et depuis des millénaires (pour ce qui est de la philosophie) a construit un cursus mature sur l'explication. Villani et al, *supra* note 25 à la p 145.

[334] L'article suivant présente plusieurs catégories de techniques basées sur ces trois critères (complexité, portée, dépendance), sachant que ces catégories ne sont ni exhaustives, ni mutuellement exclusives nous avisent les auteurs : A Adadi et M Berrada, « Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) » (2018) 6 IEEE Access 52138. Cet article offre un autre type de classification : Riccardo Guidotti et al, « A Survey of Methods for Explaining Black Box Models » (2018) arXiv:180201933 [cs], en ligne: <http://arxiv.org/abs/1802.01933>, arXiv: 1802.01933.

entre autres de la portée de ce que l'on cherche à couvrir par l'explication[335]. L'explication peut viser la compréhension d'une décision spécifique ou la compréhension du modèle de décision[336]. La portée peut rendre compte d'autres aspects d'un système, soit le processus d'apprentissage lui-même, c'est-à-dire comment le modèle est conçu; le fonctionnement de certaines portions du modèle seulement (ce qui peut être utile quand le modèle est complexe au point d'être inintelligible)[337]. Finalement, l'explication peut porter sur plusieurs cas similaires[338]. Ce qui nous intéresse ici est l'explication d'un cas spécifique, c'est-à-dire pourquoi le système a produit le résultat soumis en preuve à la cour.

5.5.2 Caractéristiques souhaitables de l'explication

L'intelligence artificielle explicable est une discipline relativement jeune qui tend à s'inspirer des sciences sociales où l'explication fait l'objet d'études depuis déjà plusieurs décennies[339]. En effet, comment choisir l'explication appropriée alors qu'il

---

[335] Molnar, *supra* note 297 ch 2.3.

[336] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You?": Explaining the Predictions of Any Classifier » (2016) KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135-1144 à la p 1135. Les auteurs associent l'intelligibilité algorithmique à la confiance envers l'algorithme et distinguent deux volets de la compréhension de l'algorithme : « It is important to differentiate between two different (but related) definitions of trust : (1) trusting a prediction, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by how much the human understands a model's behaviour, as opposed to seeing it as a black box. ». Pégny et Ibnouhsein, *supra* note 289 à la p 12. Les auteurs nomment ces distinctions « l'intelligibilité de la procédure et intelligibilité des sorties ».

[337] Molnar, *supra* note 297.

[338] *Ibid* ch 2.3.

[339] Miller, « Explanation in Artificial Intelligence », *supra* note 231.

existe une pluralité d'explications valables pour une décision ou un phénomène donné[340]? Tim Miller a révisé plus de 250 publications dans le domaine des sciences sociales afin de déterminer les points communs entre les explications de phénomènes sociaux[341]. Il constate que les caractéristiques désirables d'une explication sont les suivantes :

- L'explication est **sélective** (*selective*), c'est-à-dire qu'elle porte non pas sur l'ensemble des éléments mais les principaux éléments d'une explication, le choix des éléments étant nécessairement cognitivement biaisé.
- L'explication est **sociale**, elle transfert des connaissances entre l'émetteur et le récepteur de l'explication, permet le dialogue et l'interaction.
- L'explication est **contrastée** (*contrastive*), c'est-à-dire qu'elle répond généralement à la question *pourquoi ceci plutôt que cela*.
- L'explication est **contextualisée**. Seuls certains facteurs (associations ou causalités) sont pertinents pour le récepteur et pour l'émetteur, facteurs qui engagent l'un et l'autre dans un dialogue à l'intérieur d'un contexte précis[342]. Certains facteurs à évaluer à l'égard du contexte incluent l'impact de la décision que l'on cherche à expliquer, le coût associé à l'explication en regard de ses bénéfices, la confiance accordée au processus décisionnel et les erreurs qui pourraient s'y produire[343].

---

[340] Selbst et Barocas, *supra* note 36 à la p 1081.

[341] Miller, « Explanation in Artificial Intelligence », *supra* note 231 à la p 4.

[342] Miller, « Explanation in Artificial Intelligence », *supra* note 231.

[343] Doshi-Velez et al, « Accountability of AI Under the Law », *supra* note 288 à la section 3.

L'explication algorithmique devrait donc répondre aux caractéristiques ci-haut[344]. Elle devra également considérer :

- **Le degré de certitude** du résultat[345]. Effectivement, ce dernier peut varier d'un cas à l'autre. Par exemple, dans un cas d'identification d'un individu, un outil de reconnaissance faciale pourra accompagner le résultat d'un degré de certitude quant à l'identité trouvée. On pourrait ainsi comparer les cinq résultats présentant le plus haut degré de certitude.
- **L'importance des variables** impliquées dans la décision[346]. Selon le modèle, l'explication pourrait accorder un poids aux diverses variables impliquées ou à certaines portions de l'explication, par exemple lesquelles sont les plus importantes dans le raisonnement[347].

Le témoignage d'expert devrait fournir les principaux facteurs impliqués dans une décision ainsi que leur poids et identifier l'incertitude du résultat lorsque pertinent.

5.5.3 Exemples d'explication

Deux exemples d'explication qui pourrait être pertinente de présenter à la Cour sont l'explication contrefactuelle et l'explication contradictoire (*adversarial*). La

---

[344] Toujours selon le point de vue de l'auteur, voir Miller, « Explanation in Artificial Intelligence », *supra* note 231.

[345] Molnar, *supra* note 297 ch 3.5.

[346] *Ibid* ch 2.5.

[347] *Ibid*.

première permet d'identifier le plus petit changement qui entraînerait une décision différente de celle prise par le système[348], et la seconde permet d'identifier le plus petit changement qui entraînerait un résultat erroné[349].

5.5.3.1 L'explication contrefactuelle

L'explication contrefactuelle permet d'identifier un changement, idéalement le plus petit changement possible, qui permette d'obtenir un résultat autre que celui produit initialement par l'outil[350]. Par exemple, si l'outil d'une banque refuse un prêt, quel serait le plus petit changement dans les données qui permettrait d'obtenir ce prêt? Est-ce un revenu plus élevé, une dette moins élevée, un montant de prêt moins élevé, etc.? Le changement pertinent dépendra du contexte. Il pourrait même nous permettre d'évaluer la discrimination potentielle engendrée par un outil en analysant la variabilité des résultats selon les valeurs de caractéristiques protégées, telles que la race ou le genre[351].

Sans rentrer dans la logique interne des outils, qui peut rapidement devenir inintelligible pour des systèmes complexes qui peuvent utiliser des milliers de variables, l'explication contrefactuelle nous permet d'accéder au raisonnement qui a mené au

---

[348] *Ibid* ch 6.1.

[349] *Ibid* ch 6.2. Ce type d'explication est utile pour identifier les vulnérabilités du système à de faibles variations dans les données d'entrée.

[350] Sandra Wachter, Brent Mittelstadt et Chris Russell, « Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR » (2018) 31:2 Harvard JL & Tech 841, arXiv: 1711.00399.

[351] *Ibid* à la p 853.

résultat[352]. L'explication contrefactuelle convient parfaitement aux personnes visées par une décision car elle leur permet d'agir sur celle-ci soit en la contestant, soit en agissant sur les paramètres qui contribuent à la décision[353]. Sandra Wachter soutient que l'explication contrefactuelle répond au « droit à l'explication » tel qu'exigé par le Règlement général sur la protection des données (RGPD)[354]. Selbst et Barocas émettent cependant des réserves par rapport à ce type d'explication : ils considèrent que l'explication contrefactuelle permet essentiellement aux intéressés de naviguer plus efficacement au travers des méandres du système mais s'avère une piètre solution à la justification de la décision elle-même[355]. Ce type d'explication peut être accompagnée d'outils interactifs qui permettent d'évaluer l'incidence de variations de certaines variables sur le résultat[356]. Ces outils, qui connaissent un certain intérêt parmi les législateurs[357], offrent une certaine compréhension du raisonnement derrière la décision. Cependant leur utilité s'avère limitée lorsque le modèle est complexe et implique des interactions entre variables qui ne peuvent se traduire en règles

---

[352] *Ibid* à la p 861.

[353] Wachter, Mittelstadt et Russell, « Counterfactual Explanations without Opening the Black Box », *supra* note 351.

[354] *Ibid* à la p 863. Le « droit à l'explication » est le terme souvent utilisé dans la littérature juridique au sujet de l'explication du RGPD. CE, *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données),* [2016], JO, L119/1 [RGPD].

[355] Selbst et Barocas, *supra* note 36 à la p 1122.

[356] *Ibid* à la p 1115 : « Interactive approaches ». L'outil « What-If Tool » de Google en est un exemple, voir Sarkar, *supra* note 313.

[357] Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection Version 2.2*, 2017 aux pp 86-89.

intelligibles; ainsi ces outils peuvent donner une fausse impression de compréhension[358].

La Figure 4 ci-dessous est une image d'écran de l'outil « What-If Tool » de Google[359]. L'outil permet de modifier les valeurs des variables, par exemple l'âge d'un individu, et visualiser l'impact de ce changement sur le résultat (illustré par un point rouge ou bleu). En sélectionnant un résultat spécifique, l'outil permet d'identifier les cas les plus proches qui ont produit un résultat différent[360].

---

[358] Selbst et Barocas, *supra* note 36 à la p 1116.

[359] Google, « What-If Tool », (novembre 2021), en ligne: *Google* <https://pair-code.github.io/what-if-tool/>.

[360] Google Cloud Tech, « Using the What-If Tool for explainability », (2020), en ligne: *Youtube* <https://www.youtube.com/watch?v=jHojeFCc5HE>. L'outil offre également des fonctionnalités d'analyse de la performance globale et pour divers groupes de données.

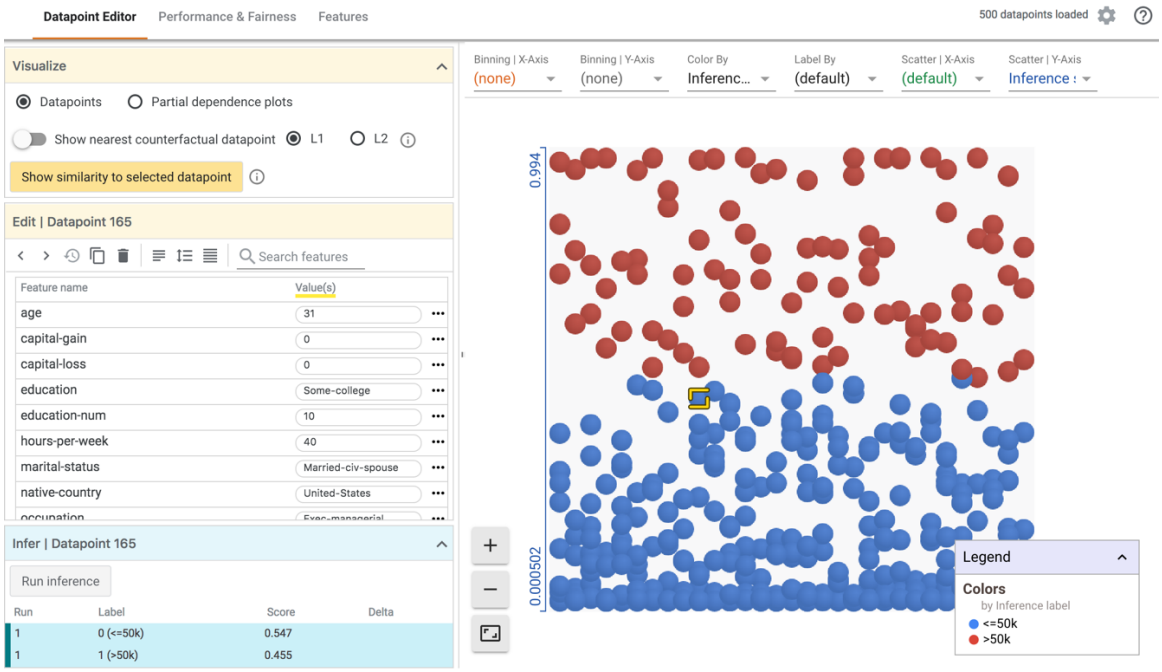

Figure 4 - Image de l'outil « What-If Tool » de Google[361]

5.5.3.2 L'explication contradictoire

L'explication contradictoire (*adversarial*) trouve son origine dans le domaine de la reconnaissance d'images où de légères perturbations dans les images, perturbations imperceptibles à l'œil nu, peuvent entraîner des erreurs dans les résultats [362] . Ainsi, un bus pourrait être incorrectement identifié comme une

[361] L'image est tiré de ce document : *Ibid.*

[362] Walt Woods, Jack Chen et Christof Teuscher, « Adversarial Explanations for Understanding Image Classification Decisions and Improved Neural Network Robustness » (2019) 1:11 Nat Mach Intell 508, arXiv: 1906.02896.

grenouille[363]! L'idée de l'explication est donc de trouver la plus petite perturbation qui entraînerait un résultat erroné. Différents exemples de perturbations pourront être fournis. Cette explication est assez similaire à l'explication contrefactuelle mais plutôt que d'identifier un changement qui produise un résultat souhaitable, elle identifie un changement qui entraîne un résultat erroné[364]. Ceci permet d'évaluer les vulnérabilités d'un système (et éventuellement de l'améliorer) et d'identifier si celles-ci pourraient se traduire en risques d'erreur pour un cas précis [365]. Les vulnérabilités des systèmes pourraient être exploitées pour berner les forces de l'ordre, par exemple. On peut penser à des filtres pour maquiller la voix et berner un système de reconnaissance vocale[366] ou au maquillage volontaire d'un objet, par exemple une arme blanche peut être maquillée pour ressembler à un parapluie afin de berner les détecteurs d'armes dans les aéroports[367].

Le témoignage d'expert devrait permettre d'engager un dialogue concernant l'impact de la modification de certains facteurs sur le résultat.

---

[363] Data Skeptic, « Adversarial Explanations », (2020), en ligne: *Youtube* <https://www.youtube.com/watch?v=L4HCAow3W3Y> autour de 00h:08m:00s.

[364] Wachter, Mittelstadt et Russell, « Counterfactual Explanations without Opening the Black Box », *supra* note 351 à la p 852.

[365] Woods, Chen et Teuscher, *supra* note 363.

[366] Khaled Khelif et al, *SIIP: An Innovative Speaker Identification Approach for Law Enforcement Agencies*, présenté à Big Data and Artificial Intelligence for Military Decision Making, 2018.

[367] Molnar, *supra* note 297 ch 10.4.

5.6 Résumé

L'objectif de l'intelligence artificielle explicable est de rendre intelligible le raisonnement par lequel le résultat d'un système d'apprentissage automatique est obtenu, et ce, en mettant en lumière les variables utilisées dans le raisonnement ainsi que leur poids. Une explication contrefactuelle ou contradictoire pourrait s'avérer pertinente pour analyser l'impact de certains changements aux données sur le résultat.

Pour jauger la fiabilité du résultat, l'explication du système d'apprentissage automatique se doit elle-même d'être fiable. Le prochain chapitre est consacré à la fiabilité de l'explication.

CHAPITRE VI

LA FIABILITÉ DE L'EXPLICATION

6.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la fiabilité de l'explication

Les questions porteront sur la fiabilité de l'explication. Elles permettent d'évaluer l'explication.

- L'explication est-elle satisfaisante?
    - Est-elle concordante avec les faits admis en preuve?
    - Rend-elle compte de l'ensemble des facteurs qui nous paraissent pertinents?
    - Quels sont les résultats les plus proches suggérés par l'outil et leur incertitude?
- L'explication décrit-elle précisément le modèle?
    - Combien de variables le modèle prend-il en compte (le modèle est-il intelligible par l'humain)?
    - Comment l'explication est-elle produite (par exemple, par une méthode *post hoc*, un modèle en soi)?
    - L'explication rend-elle compte de l'ensemble du modèle ou certaines portions?
    - Y a-t-il des explications alternatives?
    - L'explication permet-elle de prédire le résultat pour des cas distincts?
    - Deux cas similaires produisent-ils la même explication?

6.2 Expliquer et interpréter un modèle : une affaire de précision

Les termes *interprétation* et *explication* sont souvent utilisés de manière interchangeable dans la littérature. Cependant, selon Gilpin, l'interprétabilité réfère plutôt à l'intelligibilité, soit la compréhension par l'humain d'une description et ainsi elle peut être considérée comme une caractéristique de l'explication[368].

Pour la rendre intelligible, c'est-à-dire, pour que son récepteur soit en mesure de comprendre l'explication, celle-ci devra tenir compte des limites cognitives humaines[369] et des connaissances préalables des personnes à qui elle s'adressent[370]. L'intelligibilité est en en opposition avec la précision de l'explication[371]. En effet, pour rester intelligible, l'explication d'un modèle complexe devra être simplifiée et donc,

---

[368] Leilani H Gilpin et al, « Explaining Explanations: An Overview of Interpretability of Machine Learning » (2019) arXiv:180600069 [cs, stat], en ligne: <http://arxiv.org/abs/1806.00069>, arXiv: 1806.00069; Zachary C Lipton, « The Mythos of Model Interpretability » (2017) arXiv:160603490 [cs, stat], en ligne: <http://arxiv.org/abs/1606.03490>, arXiv: 1606.03490. Lipton souligne que le terme interprétabilité n'a pas de définition formelle. Par contre, l'auteur identifie les caractéristiques souhaitables d'un modèle interprétable.

[369] Notre mémoire de travail est limitée à 7 items distincts. Voir George A Miller, « The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information » (1956) 63:2 Psychological Review 81.

[370] Derek Doran, Sarah Schulz et Tarek R Besold, « What Does Explainable AI Really Mean? A New Conceptualization of Perspectives » (2017) arXiv:171000794 [cs], en ligne: <http://arxiv.org/abs/1710.00794> à la section 2. Certains modèles (« comprehensible models ») impliquent des explications qui font appel à des connaissances implicites préalables et donc leur compréhension pourra varier d'un individu à l'autre.

[371] Gilpin et al, « Explaining Explanations », *supra* note 369, art IIB. « The goal of interpretability is to describe the internals of a system in a way that is understandable to humans. The goal of completeness is to describe the operation of a system in an accurate way ». Selon l'auteur les deux caractéristiques de l'explication que sont l'interprétabilité et la complétude (« completness ») s'opposent.

perdra en précision[372]. Un modèle explicable est interprétable par défaut, mais le contraire n'est pas toujours vrai en raison, comme nous avons vu de la complexité du modèle[373].

6.3 Les modèles complexes

Plusieurs modèles d'apprentissage automatique et particulièrement les modèles d'apprentissage profond résistent à l'intelligibilité à cause de leur complexité[374]. Selbst et Barocas décrivent quatre propriétés mathématiques pour jauger la complexité des modèles[375]. Ces propriétés sont reliées à l'intelligibilité du modèle en ce sens qu'elles influencent la facilité avec laquelle nous sommes en mesure d'exécuter mentalement le modèle, c'est-à-dire de simuler les relations entre les données en entrée et le résultat

---

[372] Gilpin et al, « Explaining Explanations », *supra* note 369.

[373] *Ibid* à la p 1 : « Explainable models are interpretable by default, but the reverse is not always true ».

[374] Pégny et Ibnouhsein, *supra* note 289 à la p 9 : « Or cette capacité à comprendre de manière fondamentale les algorithmes est difficile à garantir en pratique pour les algorithmes conventionnels, et elle est à l'heure actuelle impossible à garantir pour nombre de procédures d'AM ». Note: L'auteur utilise le terme Apprentissage machine (AM) plutôt qu'apprentissage automatique; Ron Schmelzer, « Understanding Explainable AI », (23 juillet 2019), en ligne: *Forbes* <https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/> : « Many of the algorithms used for machine learning are not able to be examined after the fact to understand specifically how and why a decision has been made. This is especially true of the most popular algorithms currently in use – specifically, deep learning neural network approaches ».

[375] Selbst et Barocas, *supra* note 36. Les auteurs décrivent les quatre paramètres de complexité d'un modèle.

produit par le modèle[376]. Ces propriétés mathématiques ont trait aux relations entre les données d'entrée et le résultat en sortie :

- **Relation linéaire**. Une relation entre deux variables est linéaire lorsqu'elles varient de façon constante. Elle peut être représentée graphiquement par une droite. Un modèle non linéaire, où les variables ne varient pas de façon constante, par exemple une relation exponentielle, est plus complexe qu'un modèle linéaire. Voir les Figure 5, Figure 6 et Figure 7.
- **Relation monotone**. Une relation est monotone lorsque les variables varient dans la même direction relative. Un modèle où la variable de sortie augmente et diminue lorsque la variable d'entrée augmente est plus complexe qu'un modèle monotone. Voir la Figure 8.
- **Relation discontinue**. Une relation est discontinue lorsqu'un petit incrément d'une variable entraîne un grand bond inusité de l'autre variable. On pense à une interruption dans le tracé d'un graphique. Ce genre de modèle est complexe et plus difficile à anticiper qu'un modèle continu. Voir les Figure 9 et Figure 10.
- **Dimension**. La dimension est le nombre de variables à considérer. Les relations entre deux ou trois variables se représentent facilement sous forme de graphique, ce qui aide à comprendre les relations. Mais au-delà de trois variables, la représentation graphique n'est plus possible [377] et, outre la représentation visuelle, les subtilités des relations en jeu seront difficilement accessibles à l'humain.

---

[376] *Ibid* à la p 1096.

[377] Selbst et Barocas, *supra* note 36.

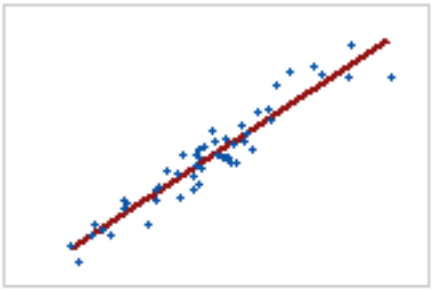

Figure 5 - Relation linéaire positive[378]

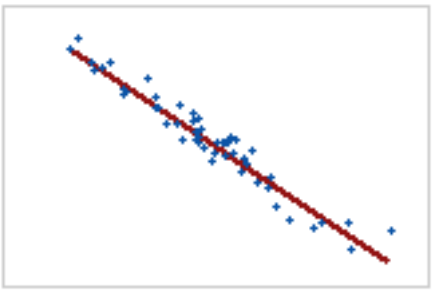

Figure 6 - Relation linéaire négative[379]

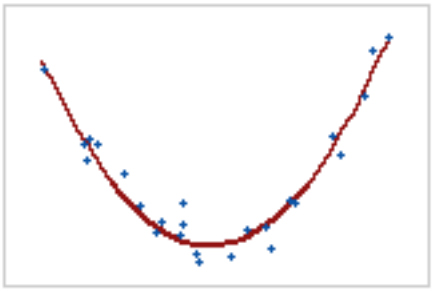

Figure 7 - Relation non-linéaire[380]

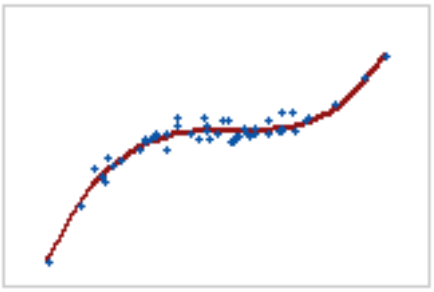

Figure 8 - Relation monotone[381]

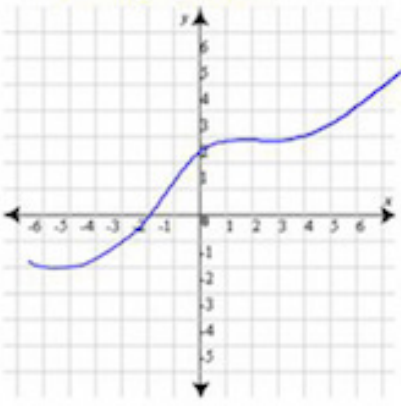

Figure 9 - Relation continue[382]

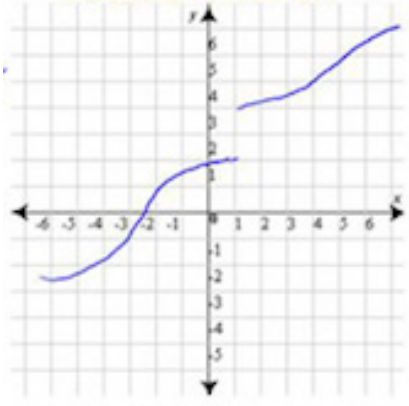

Figure 10 - Relation discontinue[383]

Outre le fait qu'un modèle puisse contenir des milliards de variables[384], même une poignée de variables dont les relations présentent les caractéristiques complexes décrites ci-haut pourrait résister à l'intelligibilité. Les modèles complexes pourront exiger des explications approximatives, tel que nous verrons dans ce qui suit.

---

[378] L'image est tirée de ce site : Minitab Statistical Software, « Diagrammes de relations linéaires, non linéaires et monotones », (2019), en ligne: *Minitab* <https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/>.

[379] L'image est tirée de ce site *ibid*.

[380] L'image est tirée de ce site *ibid*.

[381] L'image est tirée de ce site *ibid*.

[382] L'image est tirée de ce site : Studycom, « Discontinuous Functions: Properties & Examples - Video & Lesson Transcript », (novembre 2021), en ligne: *Study.com* <https://study.com/academy/lesson/discontinuous-functions-properties-examples-quiz.html>.

[383] L'image est tirée de ce site *ibid*.

[384] Marcus, « Deep Learning », *supra* note 162 : « deep learning systems have millions or even billions of parameters ».

> Le témoignage d'expert devrait indiquer le nombre de variables du modèle ou à tout le moins l'ordre de grandeur et il devrait mettre en lumière les différents types de relations entretenues entre les variables au sein du modèle afin de permettre au juge des faits de jauger la complexité du modèle.

On pourra ensuite classer les techniques d'explication algorithmique en deux grandes familles, selon la complexité du modèle sous-jacent [385].

La première famille repose sur la sélection d'un modèle intrinsèquement interprétable[386]. On pourra simplifier d'autant plus le modèle en lui imposant diverses contraintes, par exemple en limitant le nombre de variables utilisées. Un modèle tel qu'un arbre de décisions (voir la Figure 11Figure 11) entre dans cette famille d'explication : il est facilement interprétable du moment que le nombre de feuilles de l'arbre est limité, ce que l'on pourra aisément contrôler[387]. Ce type d'explication convient bien pour expliquer la logique de fonctionnement d'un système [388]. Le

---

[385] Nous utiliserons un classement basé sur la complexité du modèle. De plus, nous ne décrirons que deux familles, la 3e étant un hybride des précédentes. Adadi et Berrada, « Peeking Inside the Black-Box », *supra* note 335; Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 aux pp 22-24. La famille dite « hybride » propose d'intégrer diverses techniques lors du design même du modèle afin de limiter sa complexité et faciliter son interprétation. Par exemple, on introduira des règles de décision que le modèle optimisera lors de l'apprentissage; ces règles pourront être extraites une fois le modèle entraîné pour expliquer un résultat donné. Ces approches sont nombreuses et font l'objet de la tendance actuelle de la recherche dans le domaine de l'IA explicable.

[386] Adadi et Berrada, « Peeking Inside the Black-Box », *supra* note 335; Molnar, *supra* note 297 ch 2.2, 4.

[387] Selbst et Barocas, *supra* note 36 à la p 1111.

[388] *Ibid* à la p 1112. Le RGPD européen en est un exemple, tel que nous le verrons plus bas.

désavantage de cette technique est que l'interprétabilité inhérente du modèle se fait parfois au détriment de sa performance, de sorte que la précision des prédictions pourra en souffrir[389].

La Figure 11[390] plus bas est un exemple fictif qui illustre l'arbre de décision d'un système de prêt bancaire. La logique de la structure est la suivante : en partant du haut de l'arbre, on navigue à travers les branches selon les conditions indiquées dans les nœuds, et ce jusqu'à ce qu'on arrive à une feuille, c'est-à-dire à la décision finale quant à l'approbation ou non du prêt. Par exemple, une personne de moins de 40 ans dont les revenus sont en deçà de 166.500 et dont l'épargne mensuelle est de plus de 657 se verra accorder le prêt. L'arbre de décision dans cet exemple contient peu de variables et conditions, ce qui le rend l'explication du résultat facilement intelligible.

---

[389] *Ibid* à la p 1111.

[390] L'image est tirée de ce site : Youssef Fenjiro, « Machine learning for Banking: Loan approval use case », (7 septembre 2018), en ligne: *Medium* <https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3>.

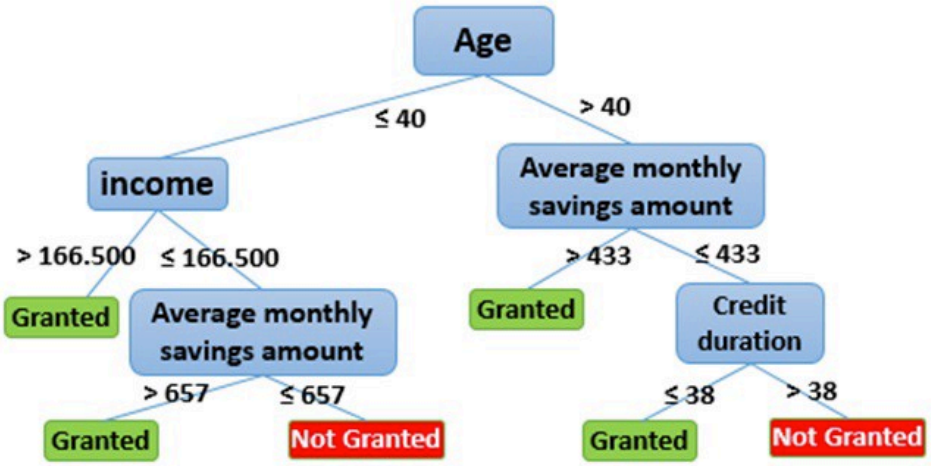

Figure 11 - Arbre de décision d'un système de prêt bancaire[391]

La deuxième famille d'explications consiste en méthodes qui découplent l'explication du modèle. L'explication est alors conçue une fois le modèle entraîné (*post hoc*)[392]. C'est une approche plus flexible que la précédente car elle n'impose pas de contraintes au modèle. Elle permet en effet de bénéficier des performances de modèles complexes et typiquement opaques tels que les réseaux de neurones[393]. Un exemple de cette approche est LIME (Local Interpretable Model-Agnostic Explanation) où un modèle simplifié est conçu par apprentissage automatique à partir d'entrées-

[391] L'image est tirée de ce site : *Ibid*.

[392] Molnar, *supra* note 297 ch 2.2, 5.

[393] Adadi et Berrada, « Peeking Inside the Black-Box », *supra* note 335.

sorties du modèle original[394]. Le modèle simplifié détecte les facteurs les plus importants impliqués dans une décision ainsi que le poids relatif de ces facteurs[395]. Ces techniques conviennent donc lorsque l'on cherche à identifier les principaux facteurs qui contribuent à un résultat spécifique[396]. LIME a l'avantage de pouvoir s'adapter à différents types de modèles, ce qui permet de comparer les explications de modèles alternatifs et sélectionner celui qui convient le mieux pour une tâche donnée[397]. Cependant, le désavantage des approches *post hoc* est qu'elles ne permettent pas d'accéder à l'ensemble des paramètres d'un modèle[398]. Elles produisent des approximations ce qui implique parfois de simplifier l'explication au risque de produire une explication trompeuse[399]. En effet, diverses approximations pourraient être en concurrence pour expliquer un même résultat[400]. Babic et ses coauteurs font référence à ce type d'explication comme un « semblant de compréhension » (*ersatz*

---

[394] Ribeiro, Singh et Guestrin, « "Why Should I Trust You? », *supra* note 299.

[395] *Ibid*.

[396] Selbst et Barocas, *supra* note 36 à la p 1114. Ce type d'explication convient bien au Equal Credit Opportunity Act américain, par exemple.

[397] Molnar, *supra* note 297 ch 5.

[398] *Ibid*.

[399] Selbst et Barocas, *supra* note 36 aux pp 1113, 1115. Un exemple d'outils interactifs est le « What-if Tool » de Google. Voir Sarkar, *supra* note 313.

[400] Boris Babic et al, « Beware explanations from AI in health care » (2021) 373:6552 Science 284.

understanding), ce qui pourrait convenir dans certains contextes si, par ailleurs, le niveau d'exactitude des résultats démontré empiriquement est satisfaisant[401].

> Le témoignage d'expert devrait préciser la méthode d'explication : est-elle conçue à l'aide d'un modèle approximatif *post hoc* ou à partir d'un modèle intrinsèquement interprétable? Si l'explication provient d'un modèle approximatif, quelles sont les diverses approximations concurrentes qui pourraient similairement expliquer le modèle et qui pourraient avoir un impact sur ce que la preuve tend à démontrer?

6.4 Critères d'évaluation de l'explication

Outre la complexité du modèle et la méthode pour la produire, on pourra considérer ces différentes caractéristiques pour évaluer l'explication algorithmique [402] :

- **La fidélité** : jusqu'à quel point l'explication est une approximation du modèle;
- **L'exactitude** : est-ce que l'explication permet de prédire un résultat à partir de nouvelles données[403] ;

---

[401] *Ibid*. Les auteurs discutent de certains contextes médicaux où parfois il est plus intéressant de se fier au résultat d'un système exact plutôt que de comprendre la logique d'un système dont l'exactitude est moindre.

[402] Molnar, *supra* note 297 ch 2.5. Selon l'auteur, il n'y a pas de consensus sur la façon d'évaluer l'explication et mesurer certaines des caractéristiques mentionnées pose des difficultés.

[403] *Ibid*. Selon l'auteur mesurer cette caractéristique est un défi.

- **La stabilité** : l'explication est-elle similaire pour deux cas semblables[404].

Dans le cadre juridique, le juge Russell Brown de la Cour Suprême, évoquait, alors qu'il était professeur de droit, certaines caractéristiques permettant d'évaluer les diverses explications soumises par les parties devant les tribunaux[405]. Nous les remettons dans le contexte de l'explication algorithmique :

- **La prise en compte des faits** (*coverage*) : l'explication prend-elle en compte les faits qui nous semblent pertinents dans le contexte?[406];
- **La cohérence** : l'explication est-elle cohérente avec les faits?[407];
- **L'unicité** : y a-t-il d'autres explications potentielles?

Les caractéristiques ci-haut permettent d'évaluer l'explication. Plus le modèle est complexe, plus il y aura d'incertitude sur l'explication et moins on pourra s'attendre à ce que celle-ci soit fiable. La valeur probante de la preuve s'en trouvera affectée[408].

---

[404] *Ibid*. Selon l'auteur mesurer cette caractéristique est un défi.

[405] Brown, « The Possibility of "Inference Causation" », *supra* note 254. Sans chercher à être exhaustif, l'auteur relève les caractéristiques suivantes : « coherence, uniqueness, completness, consistency, coverage, simplicity ». .

[406] *Ibid*. Certains faits qui nous apparaissent pertinents pourraient ne pas se retrouver dans l'explication. Il pourrait y avoir d'autres explications qui les prennent en compte. Brown rappelle qu'une explication autre que celles soumises par les deux parties au Tribunal peut être retenue par le juge des faits.

[407] Selon Tim Miller, au-delà des statistiques, ce sont surtout les relations causales ou associatives qui nous permettent de comprendre un phénomène. Voir Miller, « Explanation in Artificial Intelligence », *supra* note 231.

[408] Tel que discuté au Chapitre 1 : Limpert, « Beyond the Rule in Mohan », *supra* note 27.

Le témoignage d'expert devrait permettre de jauger l'incertitude de l'explication. Les critères d'évaluation décrits plus haut devraient être mis en lumière afin d'évaluer la fiabilité de l'explication et permettre au juge des faits d'établir le niveau de fiabilité de la preuve.

6.5 Résumé

L'explication d'un résultat produit par un système d'apprentissage automatique pourra rendre compte du fonctionnement du système, en l'occurrence des variables utilisées dans le calcul du résultat ainsi que de leur poids. Lorsqu'un modèle complexe résiste à l'intelligibilité, l'explication du résultat pourra être simplifiée, ce qui pourrait avoir pour effet d'augmenter son incertitude.

Nous nous sommes penchés sur la fiabilité d'un résultat spécifique par l'entremise de son explication. Mais qu'en est-il de la fiabilité d'un ensemble de résultats produits par le système d'apprentissage automatique? Le chapitre suivant est consacré à la performance du système qui vise à quantifier cet aspect de la fiabilité.

CHAPITRE VII

LA PERFORMANCE DU SYSTÈME

7.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la performance du système

Dans ce chapitre, nous évaluons la fiabilité de l'outil en termes quantitatifs. Comment l'outil performe-t-il sur un ensemble de cas? Contrairement aux chapitres précédents où l'on se penchait sur un résultat spécifique, ce chapitre concerne la performance globale du système. Ainsi les questions à poser sont notamment les suivantes :

- Quelles sont les métriques de performance du système dans son ensemble et pour un groupe donné (par exemple, les femmes blanches)?

- Comment se comparent la performance globale du système et la performance sur le groupe concerné dans le cas d'espèce (par exemple, une femme blanche)?

7.2 La performance globale

7.2.1 Les métriques

7.2.1.1 Les faux positifs et faux négatifs

On peut identifier deux types d'erreurs en statistiques, soit les erreurs de Type 1 qui génèrent des **faux positifs** et les erreurs de Type 2 qui génèrent des **faux négatifs** [409]. Le dénombrement des faux positifs et des faux négatifs permet d'évaluer la performance d'un système. Par exemple, supposons qu'un système tente d'identifier un chat sur une image. Il y a deux hypothèses à considérer : soit c'est un chat, soit ce n'est pas un chat. Ainsi, il y a quatre résultats possibles, tels qu'illustrés à la Figure 12, appelée *matrice de confusion*. Dans cet exemple, un système de reconnaissance d'images tente de reconnaître un chat sur 2 images, soit un chat et un plat de guacamole[410].

1- Le système détecte correctement le chat, c'est un vrai positif (VP).

2- Le système détecte que l'image n'est pas celle d'un chat alors qu'elle l'est; c'est un faux négatif (FN).

---

[409] Steven Pinker, *supra* note 225 à la session 6 « Statistical Decision Theory ».

[410] Anish Athalye et al, « Fooling Neural Networks in the Physical World », (31 octobre 2017), en ligne: *LabSix* <https://www.labsix.org/physical-objects-that-fool-neural-nets/>. L'identification erronnée d'une image de guacamole comme étant celle d'un chat est devenu un exemple classique dans le domaine de la reconnaissance d'images !

3- Le système détecte un chat alors que l'image n'est pas celle d'un chat; c'est un faux positif (FP).

4- Le système détecte correctement que l'image n'est pas celle d'un chat, c'est un vrai négatif (VN).

Dans les cas #1 (VP) et #4 (VN), le système prédit le bon résultat. Dans les cas #2 (FN) et #3 (FP), le résultat est erroné. L'événement #3 est appelé un « faux positif » ou erreur de type I : le chat est incorrectement identifié. L'événement #2 est appelé un « faux négatif » ou erreur de type II : le chat n'est pas identifié[411].

---

[411] On pourra utiliser ces métriques pour comparer et classer les performances des systèmes. Voir, par exemple, la méthodologie du NIST pour comparer les systèmes de reconnaissance faciale : National Institute of Standards and Technology, *Face Recognition Vendor Test (FRVT) Part 2: Identification, NIST IR 8271 Draft Supplement*, 2021 aux pp 11-12. Jimmy Bourque, Jean-Guy Blais et François Larose, « L'interprétation des tests d'hypothèses : p, la taille de l'effet et la puissance » (2009) 35:1 Revue des sciences de l'éducation 211. L'article précise que lors de tests d'hypothèses, la science établit que pour qu'une hypothèse soit statistiquement valide, le taux de faux positifs (erreur de type 1) doit être égal à ou inférieur à 5%. Au-dessus de 5%, une corrélation n'est plus statistiquement significative, ce qui implique qu'une hypothèse nulle (« non-chat » dans notre exemple) ne peut être rejetée et donc que l'hypothèse « chat » n'est pas concluante. Un seuil de signification statistique de 5%, signifie que 5 fois sur 100 le même résultat peut être obtenu avec l'hypothèse nulle. On peut donc conclure que l'hypothèse est statistiquement valide, ce qui ne prouve pas en soi l'hypothèse. Le seuil de signification d'un résultat doit être disponible dans toute expérience scientifique visant à confirmer ou infirmer une hypothèse. Elle représente le taux « acceptable d'erreurs » de type faux positif. De la même façon, les erreurs de type 2 (les faux négatifs) doivent aussi être quantifiées. En général, on vise 80% comme probabilité de détecter un effet s'il y a effectivement un effet (donc d'éviter les faux négatifs ou 20% d'erreur de type 2).

	CAS #1 C'est un chat (SUCCÈS) Vrai positif ✓ (VP)	CAS #2 Ce n'est pas un chat (ERREUR) Faux négatif ✗ (FN)
	CAS #3 C'est un chat (ERREUR) Faux positif ✗ (FP)	CAS #4 Ce n'est pas un chat (REJET) Vrai négatif ✓ (VN)

Figure 12 - Matrice de confusion

7.2.1.2 La relation entre faux positifs et faux négatifs

Les faux positifs et faux négatifs sont reliés par un facteur appelé « seuil de décision ». Nous l'illustrons dans ce qui suit.

Supposons que la distribution de probabilités d'une hypothèse prenne la forme d'une courbe[412], tel qu'illustré à la Figure 13. Dans cette image, la courbe mauve représente l'hypothèse du chat (h1), tandis que la courbe verte représente l'hypothèse nulle h0 (ce n'est pas un chat). Plus les courbes se superposent, plus il est difficile de discriminer le chat du non-chat.

[412] Steven Pinker, *supra* note 225 à la session 6 « Statistical Decision Theory ». On représente ici la courbe en forme de cloche, appelée « forme normale » mais elle pourrait être toute autre.

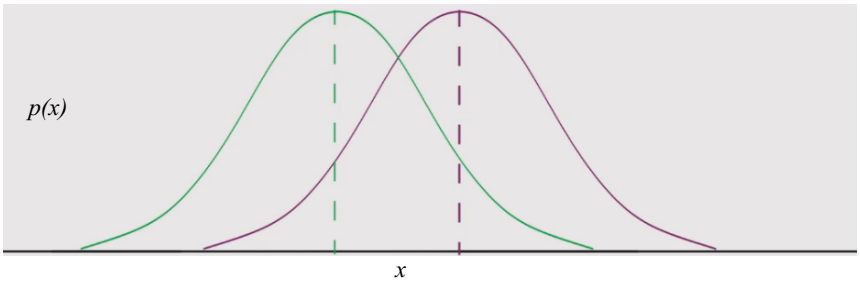

Figure 13 – Distributions de probabilités *p* d'une valeur *x*, suivant deux courbes normales[413]

La décision d'opter pour un « chat » plutôt qu'un « non-chat » dépend d'un seuil de décision, illustré par la ligne verticale rouge de la Figure 14: à droite du seuil, on optera pour la décision « chat » et à gauche du seuil, on optera pour la décision « non-chat ».

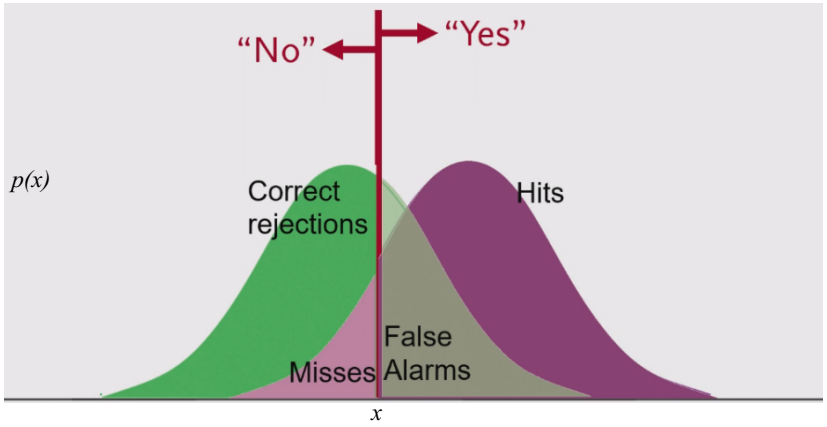

Figure 14 - Seuil de décision[414]

[413] *Ibid*. L'image est tirée de la session 6 « Statistical Decision Theory ».

[414] *Ibid*. L'image est tirée de la session 6 « Statistical Decision Theory ».

Si l'on déplace le seuil (la ligne verticale rouge) vers la droite, il y aura plus de « correct rejections » (vrais négatifs ou événement #4) mais également plus de « misses » (faux négatifs ou événement #3). Tandis que si le seuil est déplacé vers la gauche, il y aura plus de « hits » (vrais positifs ou évènement #1) mais également plus de « false alarms » (faux positifs ou événement #2). Le choix du seuil influence donc la performance du système. On pourra choisir un seuil qui augmente les succès ou augmente les rejets, avec le compromis nécessaire sur les erreurs de type faux positifs ou faux négatifs.

7.2.1.3 Métriques combinées : exactitude, précision, rappel, spécificité, F-mesure

Certaines métriques de performance combinent les faux positifs et faux négatifs, par exemple[415] :

- **Exactitude** (*accuracy)* : c'est le nombre de bonnes prédictions sur l'ensemble des cas possibles ou (VP+VN) / (VP+VN+FP+FN). Cette métrique est peu représentative lorsqu'il y a de grandes disparités entre le nombre de vrais positifs et vrais négatifs. Supposons un échantillon de 100 items où 95 items sont positifs et 5 items sont négatifs. Supposons maintenant un système prédictif complètement dysfonctionnel qui prédirait systématiquement un résultat positif. Ce système aurait une performance de 95% à cause de la seule répartition des cas dans notre échantillon, ce qui nous induirait en erreur. Les deux métriques de performance plus bas pallient ce problème.

---

[415] Les exemples sont tirés de ce site : Niwratti Kasture, « 10 Essential Ways to Evaluate Machine Learning Model Performance », (31 octobre 2020), en ligne: *Medium* <https://medium.com/analytics-vidhya/10-essential-ways-to-evaluate-machine-learning-model-performance-6bf6e11f9502>.

- **Précision** (*précision*) : la proportion de prédictions positives qui sont effectivement correctes, soit le nombre de vrais positifs (VP) divisé par l'ensemble des prédictions positives qu'elles soient vraies ou fausses ou VP / (VP+FP).

- **Rappel** (*recall*) : la proportion de prédictions positives réelles qui ont effectivement été prédites, c'est-à-dire le nombre de vrais positifs (VP) divisé par l'ensemble des positifs réels ou VP / (VP+FN).

- **Spécificité** (*specificity*) : inversement à la métrique précédente, c'est la proportion de prédictions négatives réelles qui ont effectivement été prédites ou VN / (VN+FP).

- **F-mesure** (*F1 score*) : c'est une métrique qui combine la précision et le rappel. En effet, si la précision est excellente et le rappel pauvre (ou vice versa), la F-mesure permet de considérer les deux aspects à la fois en une seule mesure.

$$F\_mesure = \frac{(2 * Précision * Rappel)}{Précision + Rappel}$$

7.2.1.4 Le choix des métriques selon le contexte

Les faux négatifs et faux positifs ont des impacts différents selon le type de problème[416]. Par exemple, diagnostiquer un cancer chez un homme alors qu'il n'en est pas atteint (faux positif) peut l'amener à subir des traitements invasifs indûment. Par contre, ne pas diagnostiquer un cancer (faux négatif) peut réduire les chances de survie et entraîner la mort. Il y a toujours un compromis à faire entre les deux types

[416] *Manuel scientifique*, *supra* note 39 à la p 98.

d'erreurs[417] et on cherchera à minimiser l'erreur la plus désavantageuse[418]. Ce compromis relève d'un choix : on voudra donc déterminer l'erreur la plus souhaitable selon le contexte[419]. Roth soutient que les forces de l'ordre auraient plutôt tendance à opter pour la réduction de faux négatifs plutôt que de faux positifs.[420]

Ainsi que l'on cherche à minimiser le nombre de faux positifs ou de faux négatifs, les métriques d'intérêt pourront varier, la précision étant plus intéressante dans le premier cas et le rappel dans le deuxième cas. Lorsque les deux objectifs importent tout autant, la F-mesure pourra être pertinente[421].

Le témoignage d'expert devrait préciser si le système à tendance à favoriser les faux positifs ou faux négatifs. La *précision* pourra être utilisée lorsqu'on cherche à minimiser les faux positifs, ce qui pourrait être le cas dans un contexte judiciaire où le résultat incrimine un suspect.

---

[417] *Ibid.*

[418] *Ibid.*

[419] Steven Pinker, *supra* note 225. Pour une excellente démonstration de ce compromis, voir la session 6 « Statistical Decision Theory ».

[420] Andrea L Roth, « Trial by Machine » (2016) 104:5 Geo LJ 48 à la p 1248.

[421] Kasture, *supra* note 416. La F-mesure est pertinente dans une distribution où les vrais négatifs sont nombreux.

7.3 Les enjeux de performance

7.3.1 La performance par sous-groupes

Malgré la bonne performance globale d'un outil, on voudra potentiellement obtenir la performance par regroupements plus granulaires car celle-ci pourrait s'avérer différente. Il pourrait y avoir d'importantes variations entre les performances obtenues sur des sous-groupes démographiques par exemple, jetant ainsi un nouvel éclairage sur la fiabilité de l'outil. Nous appellerons ce phénomène la « discrimination statistique » parce qu'il peut concerner des groupes protégés mais le même phénomène s'applique à tout autre découpage pertinent selon le contexte.

7.3.2 La discrimination statistique

7.3.2.1 Définition de la discrimination statistique

La discrimination est un concept complexe dont les multiples facettes sont sujettes à débat depuis fort longtemps, tant par les philosophes, sociologues et juristes[422]. La notion de discrimination en droit reflète bien cette complexité ; elle est vue comme étant contextuelle, relevant de la sphère culturelle au sein de laquelle se construisent des catégories sociales qui interagissent selon une dynamique évolutive[423]. De fait, non seulement la discrimination n'est-elle pas précisément définie notamment par la Constitution canadienne[424] mais le droit pourra l'envisager différemment selon

---

[422] Andrew D Selbst et al, « Fairness and Abstraction in Sociotechnical Systems » (2018) 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*) 59.

[423] *Ibid*.

[424] *Charte canadienne des droits et libertés,* partie I de la *Loi constitutionnelle de 1982*, constituant l'annexe B de la *Loi de 1982 sur le Canada* (R-U), 1982, c 11. La Charte ne définit pas la discrimination mais déclare à l'article 15 que la loi s'applique également à tous et que tous sont protégés par la loi « indépendemment de toute discrimination, notamment des discriminations fondées sur la race, l'origine

les attributs impliqués (par exemple, la religion ou l'origine ethnique), la sphère d'activité dans laquelle elle peut se déployer (par exemple, le logement ou l'emploi) et selon les acteurs impliqués (par exemple, une institution privée ou publique)[425].

À haut niveau, on peut envisager la discrimination comme étant le fait de désavantager une personne pour cause d'appartenance à un groupe social[426]. Ainsi, la discrimination s'oppose au concept d'égalité car elle implique un traitement ou une politique désavantageuse pour les membres d'un groupe par rapport à un autre groupe.

Pour les fins d'examiner les enjeux associés à la fiabilité des systèmes d'apprentissage automatique, nous allons nous concentrer ici sur une forme de discrimination statistique que l'on définira comme la surestimation ou la sous-estimation systémique des probabilités pour une population donnée[427]. Dans cette optique, la discrimination se révélera par d'importantes variations dans l'exactitude des résultats fournis par un système d'apprentissage automatique à l'intérieur de groupes démographiques. Cette forme de discrimination pourrait, selon le contexte, constituer

---

nationale ou ethnique, la couleur, la religion, le sexe, l'âge ou les déficiences mentales ou physiques ». La notion de discrimination a toutefois été abondamment définie dans la jurisprudence, notamment par la Cour suprême du Canada dans l'affaire *Withler c Canada (Procureur général)*, [2011] CSC 12 aux paras 29 et suivant. Voir généralement, concernant la notion de discrimination : Andrew Altman, « Discrimination » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2020 éd, Metaphysics Research Lab, Stanford University, 2020. La *Convention européenne pour la protection des droits humains* et le *Pacte international relatif aux droits civils et politiques* ne définissent pas plus la discrimination.

[425] Selbst et al, *supra* note 423 à la p 6.

[426] Altman, *supra* note 425 à la section 1.1 « A First Approximation ».

[427] Fjeld et al, *supra* note 264 à la p 47.

ou non de la discrimination prohibée par les instruments de droits de la personne[428]. Qu'elle le soit ou non, cette forme de discrimination affecte néanmoins l'évaluation que nous devons faire de la performance d'un système pour les groupes touchés.

7.3.2.2 Accélérateurs de discrimination : discrimination algorithmique

La discrimination, qu'on l'entende dans son sens statistique ou dans son sens juridique plus large, peut être intégrée et systématisée au sein d'algorithmes. On pourra alors parler de discrimination algorithmique. Celle-ci est pernicieuse à cause de la facilité avec laquelle on peut reproduire ces algorithmes, en multiplier la prolifération et les appliquer à grande échelle[429]. De plus, la discrimination algorithmique est plus abstraite et intangible que d'autres formes de discrimination, lesquelles se manifestent habituellement par des attitudes ou des caractéristiques plus familières et discernables[430]. Finalement, les algorithmes ont le potentiel de créer des biais insoupçonnés, du fait qu'ils reposent sur des corrélations parfois opaques entre les données[431].

---

[428] Une mesure prise pourrait être considérée comme reflétant un cas de discrimination statistique, mais ne pas constituer un cas de discrimination au sens des instruments juridiques de protection contre la discrimination. Cela pourrait notamment être le cas lorsqu'il s'agit d'une mesure découlant d'un programme d'accès à l'égalité (« affirmative action »).

[429] Xianhong Hu et al, *Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective*, UNESCO 62530, 2019 à la p 63.

[430] Sandra Wachter, Brent Mittelstadt et Chris Russell, « Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI » (2021) 41 Computer Law & Security Review, en ligne: <https://www.sciencedirect.com/science/article/pii/S0267364921000406> à la p 2.

[431] Xianhong Hu et al, *supra* note 430 à la p 63.

Le professeur Ignacio Cofone de la faculté de droit de l'Université McGill résume le potentiel discriminatoire des systèmes de décision automatisés de la sorte :

> « [ils] reflètent et accentuent les biais inhérents aux données qui leur sont fournies (les données qui servent à l'entraînement), ce qui affecte les décisions auxquelles ils aboutissent. Ils reproduisent et amplifient les scénarios nécessairement biaisés au moyen desquels ils ont été entraînés. […] La prise de décision basée sur des algorithmes peut aisément conduire à une discrimination indirecte fondée sur le genre ou la race en s'appuyant sur [des caractéristiques apparemment inoffensives] à titre d'indicateurs des traits prohibés »[432].

Le potentiel discriminatoire des algorithmes nous enjoint à examiner la performance d'un outil à partir d'un éclairage plus granulaire afin d'évaluer sa fiabilité de façon plus précise.

7.3.2.3 Exemples de discrimination algorithmique

Le National Institute Science and Technology (NIST) a réalisé une étude sur plus de 180 outils de reconnaissance faciale. L'étude a relevé une grande disparité dans la performance de ces outils sur diverses communautés, la communauté afro-américaine étant celle pour laquelle les performances étaient les moins élevées[433]. L'étude mesure la disparité des erreurs de types « faux positif »[434] et « faux négatif »[435]

---

[432] Cofone, *supra* note 302 à la section 4a.

[433] National Institute of Standards and Technology, *Face Recognition Vendor Test Part 3: Demographic Effects*, NIST IR 8280, 2019.

[434] Les faux positifs, dans ce contexte, sont des erreurs où l'outil identifie une même personne sur deux images distinctes alors que ce sont deux personnes différentes. Voir *ibid* à la p 2.

[435] Les faux négatifs sont des erreurs où l'outil ne trouve pas l'association entre deux images de la même personne. Voir *ibid*.

parmi des groupes démographiques établis selon l'âge, l'appartenance ethnique et le genre. L'étude démontre que le nombre de faux positifs est de l'ordre de 10 à 100 fois plus élevée pour des personnes d'origine africaine et asiatique que pour des personnes d'origine européenne; de même les faux positifs sont plus nombreux pour les femmes que pour les hommes. Ce déséquilibre a eu des échos critiques dans les médias en 2020 lorsqu'une erreur de type faux positif produite par un outil de reconnaissance faciale utilisé par les forces de l'ordre de la ville de Detroit a mené à l'arrestation erronée d'un homme d'origine afro-américaine[436]. Le New York Times rapporte en 2021 trois cas d'arrestation et l'emprisonnement à tort d'hommes d'origine afro-américaine aux États-Unis dans lesquels un outil de reconnaissance faciale a été mis en cause[437].

Une étude du MIT publiée en 2018[438] révèle des écarts de performance dans trois outils commerciaux de reconnaissance faciale, soit ceux d'IBM[439], Microsoft[440]

---

[436] Kashmir Hill, « Wrongfully Accused by an Algorithm », *The New York Times* (24 juin 2020), en ligne: <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>.

[437] Kashmir Hill, « Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match », *The New York Times* (29 décembre 2020), en ligne: <https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html>.

[438] Joy Buolamwini et Timnit Gebru, « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification » (2018) 81 Proceedings of Machine Learning Research 1-15. L'article est issu du projet « Gender Shades » du MIT Media Lab : Joy Buolamwini, « Gender Shades », (2021), en ligne: *MIT Media Lab* <https://www.media.mit.edu/projects/gender-shades/overview/>.

[439] IBM, « IBM Watson products », (21 juin 2021), en ligne: *IBM* <https://www.ibm.com/watson/products-services>.

[440] Microsoft, « Azure Cognitive Services », (2021), en ligne: *Microsoft* <https://azure.microsoft.com/fr-fr/services/cognitive-services/>.

et FACE++[441]. L'étude démontre que les écarts de performance entre les genres sont de l'ordre de 20% en faveur des hommes, les écarts sont de l'ordre de 20% en faveur des sujets à la peau claire versus la peau foncée. L'étude démontre la discrimination intersectionnelle, c'est-à-dire l'effet combiné de plusieurs caractéristiques discriminatoires telles que le genre et l'origine ethnique. Le pourcentage d'erreurs est de 20,8% à 34,7% pour les femmes de couleur foncée par rapport à 0% à 0,3% pour les hommes de couleur claire[442]. L'étude analyse également deux banques de données d'images avec lesquelles les tests de performance sont généralement effectués[443]. Or ces deux banques étalons sont majoritairement constituées d'hommes de couleur claire, ce qui pourrait augmenter indûment les résultats de performance globale des systèmes et limiter la détection de discrimination[444]. Ainsi, l'étude démontre l'importance de mesurer les performances sur des données standardisées qui reflètent une diversité propre à détecter la discrimination, incluant la discrimination intersectionnelle[445].

Plusieurs études soulignent ce genre de difficulté. Par exemple, une autre étude révèle des disparités semblables à celles découlant de la reconnaissance faciale, mais dans le domaine de la reconnaissance vocale[446]. Dans le domaine médical, une étude

---

[441] Face^{++}, « Face^{++} Cognitive Services », (2021), en ligne: *Face*$^{++}$ <https://www.faceplusplus.com/>.

[442] Buolamwini, *supra* note 439.

[443] Les banque de données sont IJB-A et Adience. Buolamwini et Gebru, *supra* note 439.

[444] À supposer que l'apprentissage ait été fait à partir de données similaires.

[445] Buolamwini, *supra* note 439.

[446] Allison Koenecke et al, « Racial disparities in automated speech recognition » (2020) 117:14 Proceedings of the National Academy of Sciences 7684. Dans cette étude, 5 outils de reconnaissance

publiée dans la revue Science démontre qu'un certain système d'intelligence artificielle favorise les blancs par rapport aux afro-américains parce qu'historiquement, les soins prodigués aux afro-américains sont moins onéreux que ceux prodigués aux blancs souffrant des mêmes maux, ce qui augmente le risque associé aux blancs au sein de l'algorithme et de ce fait l'attention médicale qui leur est portée[447]. Finalement, un exemple très médiatisé concerne le logiciel de recrutement de personnel d'Amazon dont les performances défavorables envers les femmes ont été reconnues par la compagnie qui a depuis banni l'utilisation de son logiciel[448].

7.3.2.4 Sources de discrimination

Les données constituent une source importante de discrimination qui se répercute au sein des algorithmes[449]. Les données reflètent les biais humains existants, ainsi le système « apprend » à reproduire ces biais[450]. Par exemple, le logiciel

---

vocale ont été évalués, soit ceux d'Amazon, Apple, Google, IBM et Microsoft. D'importantes disparités de performance défavorables à la population afro-américaine ont été relevées.

[447] Charlotte Jee, « A biased medical algorithm favored white people for health-care programs », (25 octobre 2021), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/>.

[448] Jeffrey Dastin, « Amazon scraps secret AI recruiting tool that showed bias against women », *Reuters* (10 octobre 2018), en ligne: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

[449] Fjeld et al, *supra* note 264 à la p 47; Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 à la p 44 : « Evaluation of training data is particularly important for the detection of bias ».

[450] Cofone, *supra* note 302 à la section 4a : « Les processus décisionnels automatisés reflètent et accentuent les biais inhérents aux données qui leur sont fournies (les données qui servent à l'entraînement), ce qui affecte les décisions auxquelles ils aboutissent. Ils reproduisent et amplifient les scénarios nécessairement biaisés au moyen desquels ils ont été entraînés ».

d'Amazon favorisait les hommes car ceux-ci sont majoritaires dans le domaine des technologies en général et chez Amazon en particulier[451]. Les données manquantes sont également sources de biais, par exemple les pauvres sont plus représentés au sein de services d'aide sociale que les riches, ce qui rend cette population plus sujette à diverses interventions parfois répressives de l'État, ce qui renforce d'autant plus le biais contre celle-ci[452]. Ainsi une boucle de rétroaction se met en place et amplifie le biais initial[453].

La discrimination algorithmique peut être indirecte : par exemple, des données apparemment inoffensives telles que le code postal ou la marque de voiture peuvent être recoupées pour déterminer des caractéristiques telles que l'appartenance socio-économique et même l'origine ethnique[454]. La discrimination algorithmique a également le potentiel de créer de nouvelles catégories discriminatoires autres que celles qui sont en général protégées par la loi telles que la race ou le genre[455]. Ceci est dû au fait que l'algorithme détecte des corrélations potentiellement insoupçonnées par l'humain pour créer des groupes et classifiera par la suite un individu selon sa probabilité d'appartenance à un groupe donné[456].

---

[451] Dastin, *supra* note 449.

[452] Eubanks, *supra* note 265.

[453] *Ibid*. L'autrice réfère au « feedback loop ».

[454] Monique Mann et Tobias Matzner, « Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination » (2019) 6:2 Big Data & Society 1 à la p 4.

[455] Mann et Matzner, « Challenging algorithmic profiling », *supra* note 455.

[456] *Ibid* à la p 1. Ce fonctionnement est appelé « profilage algorithmique ».

Ces particularités de la discrimination algorithmique tendent à brouiller les frontières entre les données personnelles et les données non-personnelles[457], ce qui selon Mann et Matzner amenuise les protections généralement inscrites dans la loi car celles-ci concernent les données personnelles[458]. Ainsi, les auteurs suggèrent de tenir compte de catégories plus vastes et flexibles de discrimination que celles énumérées explicitement par la loi telles que la race ou le genre[459].

Les sources de discrimination incluent également les choix, visions et présuppositions des personnes qui conçoivent et programment l'outil[460]. Par exemple, les objectifs d'apprentissage peuvent être biaisés, ceux-ci étant influencés par des décisions passées[461]. On se rappellera également que le modèle lui-même est une représentation de la réalité qui peut être biaisée[462].

---

[457] On pourra trouver une définition des « renseignements personnels » au Canada à la section Définitions : *Loi sur la protection des renseignements personnels*, LRC 1985, ch. P-21.

[458] Mann et Matzner, « Challenging algorithmic profiling », *supra* note 455. Les auteurs réfèrent au Réglement général sur la protection des données (RGPD); *RGPD*, *supra* note 355.

[459] Mann et Matzner, « Challenging algorithmic profiling », *supra* note 455 aux pp 3-4 : « a broader and more diversified approach to anti-discrimination may be an avenue to explore ». Les auteurs s'inspirent entre autres de la théorie de l'intersectionalité instiguée par l'autrice Kimberle Crenshaw en 1989 dans l'ouvrage « Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics ».

[460] Leese, Kaufmann et Egbert, *supra* note 17; Kleinberg et al, *supra* note 263.

[461] Fjeld et al, *supra* note 264 à la p 47. « The outcome of interest may be influenced by earlier decisions that are themselves biased ».

[462] Voir la section 3.3.1.2 sur les logiciels de police predictive ainsi que Leese, Kaufmann et Egbert, *supra* note 17.

7.3.3 Métriques de détection : l'effet disproportionné et l'effet disproportionné conditionnel

La discrimination statistique peut être mesurée en quantifiant les faux positifs et faux négatifs, ainsi que les études citées plus haut, par exemple l'étude du NIST sur les outils de reconnaissance faciale, ont démontré. Il existe différentes métriques qui, loin de décrire la complexité du contexte discriminatoire, demeurent néanmoins des outils pouvant être utilisés dans la conception d'algorithmes afin d'estimer la discrimination en termes mathématiques[463]. Ces métriques sont également des indices précieux interprétables par la communauté juridique afin de déceler la discrimination dans un contexte donné[464]. La communauté juridique pourra ainsi utiliser les métriques qui se rapprochent le plus de l'esprit des lois applicables.

Par exemple, selon Michael Feldman, la loi américaine sur la discrimination dans le domaine de l'emploi peut être décrite mathématiquement comme le ratio entre la proportion de résultats favorables (par exemple, une embauche) pour un groupe plus à risque selon le critère de discrimination interdite (par exemple, les femmes) et celle des résultats favorables pour le groupe moins à risque (par exemple, les hommes)[465]. C'est ce que l'on nomme « l'effet disproportionné » (*disparate impact*)[466]. Dans un tel contexte, on pourra soupçonner un effet discriminatoire si ce ratio s'avère moins de

---

[463] Selbst et al, *supra* note 423.

[464] Wachter, Mittelstadt et Russell, « Why fairness cannot be automated », *supra* note 431.

[465] Michael Feldman et al, « Certifying and removing disparate impact » (2015) arXiv:14123756 [cs, stat], en ligne: <http://arxiv.org/abs/1412.3756>, arXiv: 1412.3756. Les auteurs décrivent en termes mathématiques la règle de la « Equal Employment Opportunity Commission ».

[466] *Ibid*.

80%[467]. Bien que ce ratio ne rende pas compte de tous les aspects de la loi, par exemple des aménagements accordés par l'employeur pour accommoder des groupes protégés, il pourra néanmoins être interprété par les juristes avec la portée qui lui convient[468].

Sandra Wachter, pour sa part, relie la loi européenne antidiscriminatoire ainsi que la jurisprudence d'états européens et de la Cour de justice européenne à une métrique statistique appelée l'effet disproportionné conditionnel (*Conditional Demographic Disparity* ou *CDD*)[469]. Le CDD est une variante de la *disparité démographique*, pondérée selon un attribut donné[470]. Par exemple, dans un scénario d'embauche, il y a *disparité démographique* si pour un groupe à risque (femmes), la proportion de résultats défavorables (rejets) sur le nombre total de résultats défavorables (rejets) est supérieure à la proportion de résultats favorables (embauches) sur le nombre total de résultats favorables (embauches). Le CDD consiste à calculer la disparité démographique pour un attribut donné, par exemple le département dans lequel les embauches sont effectuées. Théoriquement, aucun attribut ne devrait produire de disparités démographiques[471].

---

[467] *Ibid.*

[468] Selbst et al, *supra* note 423.

[469] Wachter, Mittelstadt et Russell, « Why fairness cannot be automated », *supra* note 431.

[470] *Ibid.*

[471] *Ibid*. Wachter décrit en détail les ratios à la page 22. Le CDD tient compte du « paradoxe de Simpson », bien connu des statisticiens, lequel réfère au fait que l'on puisse observer une tendance dans un groupe de données et la tendance inverse dans un sous-groupe. Par exemple, on pourrait déduire à partir des statistiques globales d'admission à l'université que le processus d'admission est défavorable aux femmes, alors que si l'on analyse les statistiques par département, on réalise qu'elles sont en fait favorables aux femmes. Ce renversement de tendance lorsqu'on analyse les données par département peut s'expliquer par le fait que plus de femmes que d'hommes ont postulé auprès de départements dont le taux

Outre celles mentionnées plus haut, il existe plusieurs métriques[472] ainsi que diverses techniques[473], standards techniques[474] et outils commerciaux (dont le code est parfois libre d'accès)[475] pour prévenir, détecter et réduire la discrimination. Notons en particulier l'outil Amazon Sagemaker Clarify dans lequel la métrique du CDD a été intégrée suite à la publication de l'article de Wachter[476].

---

d'admission est très faible, ce qui explique le faible taux d'admission des femmes au total mais non par département.

[472] *Ibid*. L'article donne plusieurs exemples de mesures à la note 5 et discute en détails de la métrique appelée « negative dominance ».

[473] University of Oxford, « AI modelling tool developed by Oxford academics incorporated into Amazon anti-bias software », (avril 2021), en ligne: *Oxford Internet Institute* <https://www.oii.ox.ac.uk/news/releases/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/>; Solon Barocas, Moritz Hardt et Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019.

[474] « IEEE 7000 Projects: IEEE Ethics In Action in Autonomous and Intelligent Systems », (2021), en ligne: *IEEE* <http://ethicsinaction.ieee.org/p7000/>. Le standard IEEE P7003™ sur les biais algorithmiques est en cours de développement. ISO, « ISO/IEC TR 24027:2021 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making », (novembre 2021), en ligne: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77607.html>. Le standard ISO/IEC DTR 24027 sur les biais algorithmiques est en cours de développement.

[475] Sara A Metwalli, « 5 Tools to Detect and Eliminate Bias in Your Machine Learning Models », (2 mars 2021), en ligne: *Medium* <https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1>. Le « What-if Tool » de Google permet d'analyser la performance globale, par groupes de données et d'évaluer diverses stratégies de réduction des biais. Voir Google Cloud Tech, *supra* note 361.

[476] Stephen Zorio, « How a paper by three Oxford academics influenced AWS bias and explainability software », (1 avril 2021), en ligne: *Amazon Science* <https://www.amazon.science/latest-news/how-a-paper-by-three-oxford-academics-influenced-aws-bias-and-explainability-software>. AWS est l'acronyme de « Amazon Web Services ».

7.3.4 Le choix des métriques selon le contexte

Diverses conceptions de la discrimination peuvent parfois s'affronter, chacune menant à des résultats qui peuvent s'avérer incompatibles les uns avec les autres. Ceci est illustré de façon éloquente par une étude publiée dans le MIT Technology Review[477] sur l'outil d'évaluation de risque de récidive Compas de la compagnie Equivant (anciennement Northpointe)[478] utilisé au sein du système carcéral américain. L'étude démontre que, si historiquement les afro-américains ont été reconnus coupables de plus de cas de récidives, proportionnellement, que les caucasiens, ce qui est le cas dans plusieurs juridictions américaines où les afro-américains sont particulièrement ciblés par les forces de l'ordre[479], l'algorithme produira des résultats dont les performances sont variables entre ces deux groupes et ce en défaveur des afro-américains. Dans le cas de Compas, le pourcentage d'erreurs de type faux positif est plus élevé chez les afro-américains que les caucasiens[480]. Or, l'étude démontre qu'il est possible de régler ces écarts de performance et par le fait même tenter de rectifier le biais historique en ajustant le seuil de risque au-delà duquel l'individu est considéré un récidiviste potentiel. Mais cette solution amène d'autres problèmes, en l'occurrence il faudrait

---

[477] Karen Hao et Jonathan Stray, « Can you make AI fairer than a judge? Play our courtroom algorithm game », (17 octobre 2019), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>.

[478] L'évaluation de risque est maintenant intégrée dans la gamme de produits « Northpointe Suite » d'Equivant : Equivant, *supra* note 18. Compas n'utilise pas nécéssairement une technique d'apprentissage automatisé, par contre il classifie les individus selon des profils types auxquels sont associés une cote de risque, tout comme certains algorithmes d'apprentissage automatisé.

[479] Karen Hao et Jonathan Stray, *supra* note 478.

[480] Julia Angwin et al, « Machine Bias », (23 mai 2016), en ligne: *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

interpréter la cote de risque selon l'origine ethnique de la personne concernée[481]. Par exemple, une cote de risque de 7 sur 10 n'aurait pas le même poids si elle est attribuée à une personne d'origine caucasienne ou à une personne d'origine afro-américaine, ce qui entraînerait des conséquences distinctes pour chacune, ouvrant la porte à d'autres enjeux de discrimination[482]. En somme, puisqu'un même modèle peut défavoriser certains groupes, la solution au biais passe possiblement par deux modèles distincts, ce qui peut être inacceptable selon notre vision de l'équité[483].

Toujours concernant l'outil Compas, l'étude de ProPublica démontra que l'outil était discriminatoire envers les Afro-Américains car le pourcentage d'erreurs de type faux positif était plus élevé pour ces derniers que pour les caucasiens, ce qui dans un contexte criminel est une erreur avec des effets défavorables considérables[484]. Suite à la publication de l'étude de Propublica en 2016, la réplique de la compagnie Equivant ne se fit pas attendre : la compagnie rétorqua que le pourcentage de vrais positifs était le même pour les Afro-Américains que pour les blancs et donc, pour cette raison, l'outil

---

[481] Karen Hao et Jonathan Stray, *supra* note 478.

[482] *Ibid*.

[483] *Ibid*.

[484] Angwin et al, *supra* note 481.

n'était pas discriminatoire[485]. Qui dit vrai? En fait les deux ont raison[486]. Tout dépend de la mesure de discrimination qui nous importe le plus.

En somme, le contexte guidera le choix des métriques permettant de déceler la discrimination. Les juristes et personnes qui conçoivent les logiciels ont intérêt à collaborer afin de rendre accessible un ensemble de métriques propices à déceler la discrimination algorithmique[487]. Lorsqu'une preuve est soumise, ces métriques pourront servir de guide afin d'établir le degré de fiabilité de la preuve pour le cas précis dont il est question.

Lors du témoignage expert, les résultats des métriques de performance pour le groupe auquel appartient le suspect devraient être disponibles. On pourra choisir les métriques pertinentes en fonction du contexte judiciaire[488].

---

[485] Jacob Humerick, « Reprogramming Fairness: Affirmative Action in Algorithmic Criminal Sentencing », (15 avril 2020), en ligne: *Colum HRLR online* <hrlr.law.columbia.edu/hrlr-online/reprogramming-fairness-affirmative-action-in-algorithmic-criminal-sentencing/#post-1397-_Toc37843186>.

[486] Matthias Spielkamp, « Inspecting Algorithms for Bias », (12 juin 2017), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/>.

[487] De ce fait, Wachter préconise un dialogue continu entre les deux communautés (les juristes et les informaticiens). Voir Wachter, Mittelstadt et Russell, « Why fairness cannot be automated », *supra* note 431.

[488] Bourque, Blais et Larose, « L'interprétation des tests d'hypothèses », *supra* note 412. L'article qui est abordé à la section 7.2.1.1 du présent mémoire se penche sur les métriques généralement acceptées en science lors de tests d'hypothèses.

7.4 Résumé

Selon le contexte, certaines métriques seront plus pertinentes que d'autres pour rendre compte de la performance globale du système. Les faux positifs et faux négatifs peuvent être combinés pour évaluer diverses facettes de performance. Certaines métriques peuvent également rendre compte de disparités de performance entre sous-ensembles de données, par exemple des sous-ensembles basés sur des caractéristiques démographiques. Ces disparités pourraient potentiellement refléter une discrimination statistique.

Les mesures de performance sont obtenues à partir de tests effectués sur des ensembles de données. Pour que ces mesures soient considérées fiables, les tests et les données doivent respecter certains critères. Les tests, et de façon plus globale, l'assurance qualité, font l'objet du prochain chapitre.

CHAPITRE VIII

ASSURANCE QUALITÉ

8.1 Les questions que les juristes doivent se poser devant l'utilisation d'un système en lien avec la scientificité des tests employés

Dans ce chapitre, nous cherchons à évaluer si la méthode scientifique est bien appliquée lors des tests du système. Au-delà des tests, on évalue la conformité du système à un processus d'assurance qualité standardisé. Les questions pourraient être les suivantes :

- Quelle est la portée des tests (robustesse, sécurité)
- Les tests couvrent-ils l'ensemble du système?
- Le système est-il autonome?
- Les données de tests sont-elles variées et représentatives du cas d'espèce?
- Quelle est la taille des échantillons de tests?
- Les tests sont-ils répétables et reproductibles?
- Le système est-il conforme à des standards ou normes d'assurance qualité?
- Le système a-t-il été validé/audité par une tierce partie?

8.2 Les enjeux reliés aux tests

Nous avons vu qu'une fois le modèle entraîné, celui-ci est testé avant qu'il ne soit utilisé sur des données dans un environnement réel. Les mesures de performance vues dans le chapitre précédent sont obtenues à partir des tests effectués sur le

modèle[489]. La fiabilité des mesures de performance pourra être évaluée à partir des méthodes utilisées pour les obtenir, c'est-à-dire à partir du contexte d'exécution et paramètres des tests. Nous abordons ces caractéristiques dans les sections suivantes.

8.2.1 La représentativité des échantillons

Les tests permettent d'évaluer le modèle, en particulier en regard de deux problèmes courants en apprentissage automatique, soit le surapprentissage et sous-apprentissage (*overfitting* et *underfitting*)[490]. Dans le premier cas, le modèle est beaucoup trop spécifique et ne peut être utiliser pour généraliser et dans le deuxième cas, des corrélations potentiellement importantes sont absentes du modèle; dans un cas comme dans l'autre, le modèle ne produit pas de résultats satisfaisants sur de nouvelles entrées.

L'utilisation de données de tests suffisamment diversifiées et représentatives de cas réels facilitera la détection de ces problèmes et aura une incidence majeure sur la qualité prédictive du modèle[491]. Les données de tests doivent être validées, corrigées et enrichies au besoin[492]. Selon les bonnes pratiques, les tests devraient être effectués à plusieurs reprises sur de nouvelles données afin d'éviter, d'une part, les problèmes de

---

[489] Michael Felderer et Rudolf Ramler, « Quality Assurance for AI-based Systems: Overview and Challenges » (2021) arXiv:210205351 [cs], en ligne: <http://arxiv.org/abs/2102.05351>, arXiv: 2102.05351.

[490] Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology », *supra* note 151.

[491] Polyzotis et al, *supra* note 174. Voir également ce rapport dans le domaine spécifique de la reconnaissance faciale : Royaume-Uni, Centre for Data Ethics and Innovation (CDEI), *supra* note 162.

[492] Polyzotis et al, *supra* note 174; Royaume-Uni, Centre for Data Ethics and Innovation (CDEI), *supra* note 162.

surapprentissage[493] et, d'autre part, de tester sur des événements improbables qui peuvent parfois se produire au sein d'un même échantillon, ce qui fausserait les résultats de tests[494]. Or, comme nous l'avons vu à la section 4.2.1, le manque de données de qualité est l'un des principaux obstacles au déploiement de modèles performants[495].

> Le témoignage d'expert devrait identifier les données utilisées pour effectuer les tests, préciser la diversité des données et justifier la taille des échantillons de tests. Le juge des faits pourra s'assurer que les données de tests incluent les caractéristiques du cas d'espèce.

8.2.2 La reproductibilité

La reproductibilité réfère à la capacité de dupliquer les résultats de tests en utilisant les mêmes conditions et procédures que le test original[496]. Une étude publiée dans la revue *Nature* en 2016 fait état d'une « crise de la reproductibilité » en science[497].

---

[493] Rashidi et al, « Artificial Intelligence and Machine Learning in Pathology », *supra* note 151 à la p 13.

[494] *Manuel scientifique*, *supra* note 39 à la p 103 : « Le tribunal pourrait demander à l'expert de combien d'autres essais ou expériences mettant à l'épreuve la même hypothèse il ou elle a connaissance, et de décrire l'issue de ces études ».

[495] Notamment, ce rapport sur les outils de reconnaissance faciale souligne que la plupart des compagnies n'ont pas d'autres choix que d'acheter des modèles préentraînés, car elles n'ont pas suffisamment de données pour le faire adéquatement elles-mêmes : Royaume-Uni, Centre for Data Ethics and Innovation (CDEI), *supra* note 162 à la p 12.

[496] Bollen et al, *supra* note 101.

[497] Monya Baker, « 1,500 Scientists Lift the Lid on Reproducibility » (2016) 533:7604 Nature 452.

Plus de 70% des 1576 chercheurs interviewés rapportent avoir été incapables de reproduire l'expérience d'un pair. L'intelligence artificielle n'est pas à l'abri de ce constat[498]. En plus de ce phénomène général, la reproductibilité des tests d'outils d'apprentissage automatique pose des défis spécifiques à cette technique[499]. En effet, un système d'apprentissage comporte plusieurs paramètres intrinsèquement aléatoires qui complexifient la reproduction des tests[500]. Par exemple, l'échantillonnage aléatoire au départ d'un test[501], les paramètres utilisés dans la comparaison entre le résultat et la prédiction[502] ainsi que la parallélisation des calculs[503], peuvent engendrer des résultats distincts d'un essai à l'autre. Parmi les autres freins à la reproductibilité, notons l'infrastructure colossale parfois requise pour effectuer les tests et que peu d'institutions peuvent se procurer ainsi que les enjeux de propriété intellectuelle du

---

[498] Odd Erik Gundersen et Sigbjørn Kjensmo, « State of the Art: Reproducibility in Artificial Intelligence » dans *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. L'auteur rapporte que sur 400 publications spécialisées en intelligence artificielle, aucune ne contient l'ensemble des facteurs permettant de reproduire l'expérience.

[499] La reproductibilité est l'une des trois caractéristiques d'un résultat scientifique fiable selon les auteurs : « We view robust scientific findings as ones that are reproducible, replicable, and generalizable. ». Voir Bollen et al, *supra* note 101.

[500] Peter Sugimura et Florian Hartl, « Building a Reproducible Machine Learning Pipeline » (2018) arXiv:181004570 [cs, stat], en ligne: <http://arxiv.org/abs/1810.04570>, arXiv: 1810.04570; Gundersen et Kjensmo, *supra* note 484.

[501] Prabhat Nagarajan, Garrett Warnell et Peter Stone, « The Impact of Nondeterminism on Reproducibility in Deep Reinforcement Learning » (2018) International Conference on Machine Learning, en ligne: <https://openreview.net/forum?id=S1e-OsZ4e7>.

[502] Jennifer Villa et Yoav Zimmerman, « Reproducibility in ML: why it matters and how to achieve it », (25 mai 2018), en ligne: *Determined AI* <https://determined.ai/blog/reproducibility-in-ml/>.

[503] *Ibid*.

code et des données[504]. Les solutions au problème de reproductibilité incluent la publication du code source[505] et l'utilisation dès la conception des systèmes de méthodes et outils favorisant la reproductibilité[506].

8.2.3 La répétabilité (*replicability*).

La répétabilité est la capacité à dupliquer les résultats de tests en utilisant les mêmes conditions et procédures mais de nouvelles données[507]. Le fait que le modèle puisse évoluer complexifie la répétabilité d'un test, l'une des caractéristiques de la robustesse d'un test[508]. Selon un article de l'AFCEA, une ONG qui traite du domaine militaire, les tests sont difficiles à répéter lorsque les paramètres des algorithmes d'apprentissage automatique s'ajustent au fur et à mesure qu'ils traitent des données[509].

---

[504] Will Douglas Heaven, « AI is wrestling with a replication crisis », (12 novembre 2020), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/>.

[505] Elizabeth Gibney, « This AI researcher is trying to ward off a reproducibility crisis » (2019) 577:7788 Nature 14.

[506] Sugimura et Hartl, *supra* note 501; Heaven, *supra* note 505. Joelle Pineau, chercheure à l'Université McGill, contribue à rendre l'IA reproductible, par exemple en établissant la liste d'éléments que les chercheurs se doivent de fournir pour rendre leurs recherches reproductibles.

[507] Bollen et al, *supra* note 101.

[508] La répétabilité est l'une des trois caractéristiques d'un résultat scientifique fiable selon les auteurs: « We view robust scientific findings as ones that are reproducible, replicable, and generalizable. ». Voir *ibid* aux pp 2-3.

[509] Capt Jason Nunes, « DOD Policy Ignores Machine Learning », (16 mars 2020), en ligne: *Signal Magazine* <https://www.afcea.org/content/dod-policy-ignores-machine-learning>. Ces systèmes autonomes sont traités à la prochaine section.

L'article propose de réviser les standards de tests pour accommoder les particularités de l'apprentissage automatique[510].

Le témoignage d'expert devrait confirmer que les tests sont reproductibles et répétables ou, dans le cas contraire, préciser l'impact sur l'incertitude quant aux mesures de performance du système.

8.2.4 Autonomie et opacité des modèles

Plusieurs systèmes d'apprentissage automatique sont conçus pour évoluer de façon autonome de façon à s'adapter automatiquement à de nouvelles données. Les modifications des logiciels traditionnels sont toujours testées avant d'être utilisées dans un environnement réel. Dans le cas de systèmes qui évoluent en production, le processus de tests est complexifié[511].

Le modèle final n'est pas connu à l'avance et une fois entraîné, ce dernier peut rester difficile à interpréter, voire insondable. Les méthodes traditionnelles de tests qui se basent sur les spécifications précises de l'algorithme ne peuvent donc pas être appliquées. Il est donc à la fois difficile de tester un modèle ou même de confirmer l'étendue des tests, faute de comprendre le modèle[512].

---

[510] *Ibid*.

[511] Felderer et Ramler, « Quality Assurance for AI-based Systems », *supra* note 490.

[512] *Ibid*.

Si le système est autonome, le témoignage d'expert devrait le préciser et indiquer à quel moment ont été effectués les tests.

8.3 Assurance qualité

Jusqu'à présent nous avons abordé les tests du modèle, une fois ce dernier entraîné et en particulier les tests de performance. Cependant, les tests de performance ne couvrent pas l'ensemble des fonctionnalités du système. De plus, ils ne permettent que de détecter les erreurs une fois le modèle produit, ce qui peut inciter à corriger les symptômes plutôt que la source du problème car reprendre des étapes ultérieures de conception et développement peut s'avérer coûteux. En ce sens, les résultats tirés d'analyses réalisées à l'aide d'un système qui fait l'objet d'un processus d'assurance qualité adéquat seront plus fiables que celles tirées d'un système semblable mais qui ne jouirait pas d'un tel processus de contrôle.

La norme ISO/IEC 25000:2005 sur l'ingénierie logicielle définit le processus d'assurance qualité comme suit : « the systematic examination of the extent to which a software product is capable of satisfying stated and implied needs »[513]. Cette caractéristique d'un système de « faire ce pourquoi il est conçu » (*perform as intended*) fait partie intégrante de la fiabilité d'un système[514]. L'assurance qualité est un

---

[513] *Ibid* à la p 2 . Les auteurs citent la norme ISO/IEC 25000:2005 « Ingénierie du logiciel - Exigence de qualité du produit logiciel et évaluation » à la note 3.

[514] Fjeld et al, *supra* note 264 à la p 37 : « A system that is reliable is safe, in that it performs as intended, and also secure, in that it is not vulnerable to being compromised by unauthorized third parties ».

processus en continu qui a lieu tout au long du cycle de vie du système, à partir de sa conception et tout au long de son opérationnalisation[515]. Le processus permet d'améliorer le système. Ainsi la correction d'erreurs ou les leçons tirées de la rétroaction humaine sont continuellement réinjectées dans ce dernier pour l'améliorer ou l'ajuster aux conditions changeantes de son environnement[516]. Le processus d'assurance qualité des logiciels revêt plusieurs dimensions, notamment[517] :

- La robustesse, c'est-à-dire la résilience par rapport aux perturbations de l'environnement;
- La sécurité, c'est-à-dire la résilience par rapport à menaces externes, par exemple l'accès illégal aux composantes par une tierce partie[518].

Le témoignage expert devrait indiquer la portée des tests qui, en plus des tests de performance, devrait couvrir la robustesse et la sécurité du système.

---

[515] Felderer et Ramler, « Quality Assurance for AI-based Systems », *supra* note 490. L'article distingue le « offline testing » et « online testing », le premier a trait aux tests lors du développement et le second aux tests lorsque le système est opérationnel. Roth, *supra* note 23. Dans un contexte légal, l'autrice met en évidence la nécessité d'un examen minutieux des processus et procédures de manipulation des données lors de la conception et du développement de système.

[516] Fjeld et al, *supra* note 264 à la p 31.

[517] Felderer et Ramler, « Quality Assurance for AI-based Systems », *supra* note 490.

[518] Fjeld et al, *supra* note 264 à la p 39 : « The principle of "security" concerns an AI system's ability to resist external threats ». CE, Commission, *Lignes directrices en matière d'éthique pour une IA digne de confiance: Groupe d'experts de haut niveau sur l'intelligence artificielle*, 2019 à la p 20. Le projet de lignes directrices en 2018 mentionne au par. 11 que la sécurité revêt une importance particulière pour les systèmes d'IA utilisés par les forces de l'ordre et les services judiciaires. Ainsi les systèmes « doivent être soigneusement examinés et suffisamment solides et résilients pour prévenir les conséquences potentiellement catastrophiques d'attaques malveillantes ». Dans la version de 2019, les lignes directrices soulignent la prise en compte de la « résilience aux attaques et sécurité » face à l'utilisation potentiellement abusive par des acteurs malveillants à la p. 20.

8.3.1 Standards et normes

Il existe des standards techniques en assurance qualité spécifiques à l'intelligence artificielle et l'apprentissage automatique, par exemple les standards ISO et IEEE[519]. Les standards pourront couvrir l'ensemble des aspects de la fiabilité discutés dans cet ouvrage[520] et ce, durant toute la durée de vie du système[521]. Notons que certains standards prennent en considération les exigences réglementaires, tels que IEEE[522]. Ainsi, la conformité des systèmes d'apprentissage automatique aux exigences d'un corpus juridique pourrait être facilitée par le biais de mécanismes de certification à divers standards.

8.3.2 Revue et audits

L'un des principaux outils de l'assurance qualité est l'audit ou revue de code, de documentation ou tout autre artefact du cycle de vie logiciel[523]. L'audit pourra être réalisé par une partie indépendante, tel que préconisé par plusieurs lois, comme nous

---

[519] ISO, *supra* note 475; note 475.

[520] Voir, par exemple, le standard The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2019 aux pp 17-18.

[521] Kishan Milan, « Quality Assurance Factors: Key Aspects for Software Quality Control and Testing » (2021) 69 International Journal of Computer Trends and Technology.

[522] *IEEE Global Initiative*, *supra* note 521 à la p 217. L'initiative IEEE, par exemple, se penche sur les interactions entre le droit et les standards.

[523] Milan, « Quality Assurance Factors », *supra* note 522.

le verrons au chapitre 9[524]. Pour ce faire, les artefacts doivent être accessibles, ce qui pourra poser des enjeux sur le plan du secret commercial.

8.3.2.1 L'exemple du génotypage probabiliste aux États-Unis : l'impact des audits

L'exemple du *génotypage probabiliste* illustre bien les bénéfices de l'audit en ce qui concerne les décisions judiciaires. Les outils d'analyse d'échantillon d'ADN simple, provenant d'un ou deux individus, sont réputés fiables[525]. Au Canada, ceux-ci sont généralement admissibles en preuve sur la base de leur fiabilité[526]. Cependant, une génération ultérieure d'outils appelée *génotypage probabiliste* (« probabilistic genotyping ») [527] permet aujourd'hui d'analyser des échantillons complexes d'ADN provenant de plusieurs personnes ou dont les quantités sont infimes[528]. Des exemples

---

[524] Par exemple, l'exigence 6.2.5.3 de la Directive canadienne : Canada, Conseil du Trésor, *Directive sur la prise de décision automatisée*, 2019.

[525] États-Unis, President's Council of Advisors on Science and Technology*, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods,* 2016 [President's Council of Advisors] à la p 147.

[526] *Trochym*, *supra* note 65 au para 32 : « Bien que certaines formes de preuve scientifique gagnent en fiabilité avec le temps, d'autres peuvent reculer en raison de problèmes mis au jour par de nouvelles études. Ainsi, une technique qui a déjà été admissible peut un jour être jugée inadmissible. À titre d'exemple de technique scientifique qui, après s'être raffinée et avoir été étudiée de façon plus approfondie, est devenue suffisamment fiable pour être utilisée dans les procès criminels, citons l'analyse de l'empreinte génétique, que la Cour a reconnue dans R. c. Terceira, [1999] 3 R.C.S. 866 ». Nicole Duval Hesler, « L'admissibilité des nouvelles théories scientifiques » (2002) 62 R du B 359. Les résultats sont le plus souvent contestés par le truchement de la fiabilité des échantillons plutôt que sur la fiabilité des outils.

[527] Presser et Robertson considèrent le génotypage probabiliste comme de l'intelligence artificielle tandis que Richmond semble les distinguer tout en faisant ressortir leurs similarités. Presser et Robertson, *supra* note 115; Karen McGregor Richmond, « AI, Machine Learning, and International Criminal Investigations: The Lessons From Forensic Science » (2020) iCourts Working Paper Series, No 22, en ligne: <https://papers.ssrn.com/abstract=3727899>.

[528] Richmond, « AI, Machine Learning, and International Criminal Investigations », *supra* note 528.

de tels outils sont TrueAllelle, Forensic Statistical Tool (FST) et STRmix[529]. Le génotypage probabiliste est couramment utilisé par les forces de l'ordre et généralement admis en preuve aux États-Unis[530]. Au Canada, bien que leur utilisation ne cesse de croître, leur admissibilité n'a pas fait l'objet de débats au sein des tribunaux selon la Commission du droit de l'Ontario[531].

Or suite à la divulgation du code source de FST et STRmix, plusieurs erreurs ont été relevées[532], ce qui fit dire à la Cour du New Jersey dans l'affaire *State of New Jersey v Corey Pickett* que FST, utilisé dans des milliers d'affaires devant les tribunaux, n'était pas fiable, ne fonctionnait pas et se devait d'être éliminé[533]. Les erreurs détectées dans STRmix auraient quant à elles impacté une douzaine de cas aux États-Unis[534] et selon un journal australien soixante cas en Australie[535].

---

[529] *President's Council of Advisors*, *supra* note 526 à la p 78. Cybergenetics, « Cybergenetics complements the crime laboratory, making more science happen », (2021), en ligne: *Cybergenetics* <https://www.cybgen.com/services/crime_lab_complementor/page.shtml>.

[530] Roth, *supra* note 23 n 247.

[531] Presser et Robertson, *supra* note 115 à la p 15. Le rapport fait état de cinq cas au Canada où il est question d'outils de génotypage probabiliste. Selon le rapport, l'admissibilité de ce type de preuve n'a pas été contestée. Voir notamment *R v Dosanjh*, [2019] 2019 ONSC 1320 .

[532] *State of New Jersey v Corey Pickett*, [2021] 466 NJ Super 270, 2021 NJ Super Lexis 17 à la p 6.

[533] *Ibid* : « [the source code] review demonstrated the software—employed in thousands of criminal prosecutions—was unreliable, did not work as intended, and had to be eliminated ».

[534] Katherine Kwong, « The Algorithm says you did it: the use of black box algorithms to analyze complex DNA evidence. » (2017) 31:1 Harvard Journal of Law & Technology 275 à la p 292.

[535] David Murray, « Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases », (20 mars 2015), en ligne: *The Courier Mail*

Un autre exemple concerne l'affaire *People vs Hillary*[536] où l'analyse d'ADN a été réalisée par deux outils, TrueAllelle et STRmix. Or, les deux outils produirent des résultats distincts[537]. On pourrait conclure à l'instar du *President's Council of Advisors on Science and Technology* que même si les méthodes utilisées par les outils sont scientifiquement valides, l'implantation de ces méthodes par l'outil peut être erronée[538]. Une évaluation du code pourrait permettre de déceler ces erreurs. Il est inquiétant que deux outils produisent des résultats distincts car les parties en cause dans une affaire criminelle pourraient tester divers outils jusqu'à l'obtention du résultat souhaité[539].

Le témoignage expert devrait indiquer si le système est conforme à des standards reconnus, s'il est certifié ou a été audité. Les résultats produits par des techniques ou outils concurrents, pourraient affecter la valeur probante de la preuve.

<https://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b>.

[536] Roth, *supra* note 23 à la p 2020. L'auteure réfère à cette affaire à la note 254.

[537] *Ibid*. Le juge a finalement refusé les résultats provenant de STRmix pour non-conformité aux normes applicables de la part du laboratoire qui a fourni les échantillons.

[538] *President's Council of Advisors*, *supra* note 526 à la p 79.

[539] Katherine Kwong, « Harv JL & Tech », *supra* note 535 à la p 294.

8.4 Résumé

Dans ce chapitre, nous avons identifié certains critères de la méthode scientifique concernant les tests: la reproductibilité et la répétabilité des tests ainsi que la représentativité des données de tests. Au-delà des tests de performance, la qualité de l'ensemble des étapes de la conception à l'opérationnalisation du système ainsi que de l'ensemble de ses facettes, notamment sa résilience aux changements ou aux attaques externes, peut servir d'indicateurs de fiabilité du système. La conformité du système à une norme de qualité, par exemple ISO ou IEEE, de même que les audits et revues par les pairs sont autant d'outils permettant d'évaluer la qualité d'un système.

Tout au long de ce mémoire, nous avons identifié diverses facettes de la fiabilité d'un système d'apprentissage automatique. Quels sont les instruments réglementaires qui appuient et complètent cette démarche? Le chapitre suivant est consacré à ce sujet.

CHAPITRE IX

LA RÉGLEMENTATION

9.1 Support réglementaire à l'évaluation de la fiabilité

La nécessité d'une réglementation spécifique pour faire face aux systèmes automatisés utilisant les techniques d'intelligence artificielle fait largement consensus, comme en témoignent les recommandations de membres de la communauté juridique, académique, de la société civile, de l'industrie et d'institutions internationales, et ce dans divers contextes d'utilisation[540]. Malgré ces appels aux législateurs, peu de gouvernements ont adopté des lois spécifiques sur l'intelligence artificielle[541], l'Union

---

[540] Québec, Commission des droits de la personne et des droits de la jeunesse, *Mémoire à la Commission d'accès à l'information sur le document de consultation «intelligence artificielle»*, 2020 : « Même si le cadre juridique que constitue la Charte, qui, rappelons-le, a une valeur quasi constitutionnelle, s'applique déjà à l'utilisation des nouvelles technologies, il apparaît utile de l'encadrer plus spécifiquement de façon à parer aux nouvelles formes que pourraient prendre les atteintes aux droits et libertés ». Xianhong Hu et al, *supra* note 430 à la p 28 : « Develop norms and policies for improving openness, transparen-cy and accountability in automated decisions taken by AI systems through methods such as ex-ante information disclosure and ex-post monitoring of automated decision-making ». Instruments juridiques OCDE, *Recommandation du Conseil sur l'intelligence artificielle,* no de doc OECD/LEGAL/0449, 2019 à la section 2 « Politiques nationales et coopération internationale à l'appui d'une IA digne de confiance ». Partnership on AI, *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System,* 2019 [Partnership on AI].

[541] Kathleen Walch, « AI Laws Are Coming », (20 février 2020), en ligne: *Forbes* <https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/> : « It may not be surprising to find that most governments are adopting a "wait and see" approach to laws and regulations

européenne étant particulièrement active dans ce domaine et les États-Unis plutôt passifs[542]. Nous abordons dans ce chapitre les aspects réglementaires qui touchent aux concepts de fiabilité couverts dans cet ouvrage, l'objectif étant de donner des outils pour faciliter l'évaluation de la fiabilité d'une preuve issue de l'apprentissage automatique. Nous ne nous limiterons pas au Canada puisque d'autres législations peuvent nous inspirer, orienter notre réflexion et justifier nos questions adressées à l'expert sur la fiabilité d'un outil lors d'un témoignage au tribunal. De plus, nous ne nous limiterons pas aux lois en vigueur car plusieurs recommandations de la part de juristes ou projets de lois nous paraissent pertinents pour ces mêmes raisons.

9.2 Recommandations de la communauté juridique

Les lois ou projets dont nous discutons plus bas reconnaissent généralement le potentiel discriminatoire des algorithmes et leur opacité. Ils exigent en général une explication du résultat produit par le système, explication qui permette la transparence des données personnelles utilisées dans le système et la possibilité de les corriger en cas d'erreurs[543].

---

on AI. Just like with any new technological wave it's hard to predict just how this new technology will be used, or abused ».

[542] *Ibid*. L'article porte sur un rapport concernant l'état des lois sur l'IA dans le monde (Worldwide AI Laws and Regulations par Cognilytica) : « The European Union is the most active in proposing new rules and regulations [...]. On the other hand, the United States maintains a "light" regulatory posture when it comes to laws around AI ».

[543] Selbst et Barocas, *supra* note 36 à la p 1091. Voir la note 32 : « There must be no personal-data record-keeping systems whose very existence is secret ».

Toutefois, la communauté juridique a émis plusieurs recommandations afin d'élargir la portée de la règlementation pour inclure les méthodes de conception et développement de systèmes. Par exemple, Selbst et Barocas proposent d'accéder à la documentation permettant de justifier l'ensemble des décisions prises en cours de conception du système[544]. Ceci permettrait de comprendre *pourquoi* le système produit un résultat spécifique plutôt qu'uniquement *comment* ce dernier est produit, aspect sur lequel les lois insistent en général, selon les auteurs[545]. De même, Roth suggère de mettre la lumière sur l'ensemble des procédures de conception et développement des systèmes car elles sont guidées par des choix que l'on doit rendre transparents[546]. Pour Kluttz et Eaglin la conception et le développement de systèmes sont guidés par une vision et des orientations spécifiques qui se doivent d'être transparentes[547].

Nous verrons plus bas que deux instruments issus des travaux de la Commission européenne sont arrimés à ces recommandations, soit la *Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union*[548] et la *Résolution du Parlement européen adoptée le 6*

---

[544] *Ibid* à la p 1117 : « Notably, neither the techniques nor the laws go beyond describing the operation of the model. Though they may help to explain why a decision was reached or how decisions are made, they cannot address why decisions happen to be made that way ».

[545] *Ibid* à la p 1130.

[546] Roth, *supra* note 23.

[547] Kluttz, Kohli, Nitin, et Mulligan, Deirdre K., *supra* note 307; Jessica Eaglin, « Constructing Recidivism Risk » (2017) 67:59 Emory LJ 60.

[548] CE, Commission, *Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union* 2021/0106 (COD), 2021.

octobre 2021 portant sur l'utilisation de l'IA par les autorités policières et judiciaires dans les affaires pénales[549]. Ce dernier instrument est particulièrement pertinent dans le cadre de ce mémoire car sa portée concerne spécifiquement les affaires pénales[550].

Le reste du chapitre est donc consacré à divers instruments juridiques qui abordent la fiabilité des systèmes d'intelligence artificielle et dont la portée est variable. Nous nous intéressons en premier lieu aux instruments européens, notamment le RGPD qui fait office de précurseur dans la réglementation de l'intelligence artificielle. Nous abordons deux instruments juridiques français, le premier précise les éléments requis pour expliquer le processus de décision et le second pose la question d'un traitement potentiellement différencié pour les systèmes autonomes[551]. Finalement, la résolution et la proposition du Parlement européen évoquées plus haut sont abordées car elles couvrent la majorité sinon l'ensemble des notions de fiabilité abordées dans ce mémoire.

Nous enchaînerons avec les États-Unis. Le contraste entre les deux approches, européenne et américaine, nous permettra d'apprécier le spectre des exigences réglementaires, les unes étant plus étoffées que les autres. Nous présenterons finalement la réglementation au Canada et au Québec, que l'on pourrait situer à mi-chemin dans ce spectre.

---

[549] *Résolution sur l'IA en droit pénal*, *supra* note 186.

[550] Cet instrument suscite une réflexion sur le niveau de granularité requis par la portée réglementaire.

[551] Selon l'interprétation suggérée par Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308.

9.3 Europe

9.3.1 Le Règlement général sur la protection des données (RGPD)

En Europe plusieurs instruments s'appliquent à l'intelligence artificielle[552]. En l'occurrence, le Règlement général sur la protection des données (RGPD)[553], réputé précurseur de la règlementation de l'intelligence artificielle [554] , s'applique à l'administration publique de même qu'à l'entreprise privée[555]. Le RGPD reconnaît notamment :

- Le potentiel discriminatoire de la prise de décisions basée sur des algorithmes[556];

---

[552] Dr Gloria Gonzalez Fuster, *Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights*, Département thématique des droits des citoyens et des affaires constitutionnelles du Parlement européen, 2020 à la section 2. Par exemple, la Directive « Police Justice », tel qu'indiqué à l'article 1 vise : « la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, y compris la protection contre les menaces pour la sécurité publique et la prévention de telles menaces » : CE, *Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil*, [2016], JO, L119/89.

[553] *RGPD*, *supra* note 355.

[554] Noel Corriveau, « Regulating automated decision systems in Canada: What it means for your business », (23 décembre 2020), en ligne: *INQ Law* <https://inqdatalaw.medium.com/regulating-automated-decision-systems-in-canada-what-it-means-for-your-business-bdbb04d6c725>.

[555] Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 à la p 32.

[556] *RGPD*, *supra* note 355 au para 71.

- Le droit à l'explication[557];
- L'obligation du responsable du traitement de corriger les erreurs dans les données personnelles[558].

En effet, le paragraphe 71 du préambule y prévoit que le responsable du traitement algorithmique doit « sécuriser les données à caractère personnel d'une manière qui […] prévienne, entre autres, les effets discriminatoires à l'égard des personnes physiques[559] et « faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le risque d'erreur soit réduit au minimum »[560].

Un traitement automatisé qui utilise des données personnelles doit être assorti d'une explication[561]. Toute information et communication relatives au traitement de ces données doivent être « aisément accessibles, faciles à comprendre, et formulées en des termes clairs et simples »[562]. Une personne concernée par un traitement automatisé est en droit de recevoir, notamment « l'accès aux données à caractère personnel », « la

---

[557] *RGPD*, *supra* note 355.

[558] *Ibid* au para 71.

[559] *Ibid*.

[560] *Ibid*.

[561] *RGPD*, *supra* note 355.

[562] *Ibid* au para 39.

source d'où proviennent [ces données] » ainsi que « des informations utiles concernant la logique sous-jacente [du traitement] »[563].

Le RGPD, sans préciser les détails du contenu de l'explication ou son format[564], exige donc que les données utilisées et la logique du traitement puissent être communiquées à la personne concernée de façon à ce que celle-ci comprenne les raisons d'une décision[565].

Le RGPD fait l'objet de vifs débats entre acteurs du milieu juridique en ce qui concerne l'interprétation du « droit à l'explication »[566]. Parmi ses détracteurs, Wachter soumet que le « droit à l'explication » n'est pas légalement contraignant[567]. L'auteure

---

[563] *Ibid*, arts 13-15.

[564] Deeks, *supra* note 286 à la p 1849 : « it remains unclear precisely what the GDPR [RGPD] requires and what steps states and companies must take to meet those requirements ». Selbst et Barocas, *supra* note 36. Le RGDP fait l'objet de nombreux débats sur l'interprétation du « droit à l'explication ». Selon les auteurs, certains reconnaissent un droit explicite à l'explication et d'autres pas.

[565] France, Groupe de travail « article 29 » sur la protection des données, *Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679*, 2018 à la p 28. Le responsable du traitement n'a pas à fournir une explication complexe ou divulger l'algorithm complet. « Les informations fournies doivent toutefois être suffisamment complète pour que la personne concernée comprenne les raisons de la décisions ». À la note 40 : « La complexité ne peut excuser l'absence de fourniture d'informations à la personne concernée ».

[566] Selbst et Barocas, *supra* note 36 n 143. Le « droit à l'explication » fait l'objet de nombreux débats. Les auteurs réfèrent à plusieurs auteurs dont les positions divergent.

[567] S Wachter, B Mittelstadt et L Floridi, « Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation » (2017) 7:2 International Data Privacy Law 76.

relève le manque de précision du RGPD qui n'énonce ni les objectifs de l'explication, ni son contenu, ni sa portée exacte[568].

9.3.2 La Loi pour une république numérique en France

En France, la *Loi pour une république numérique*[569] qui modifie le *Code des relations entre le public et l'administration*[570] exige, à propos d'un traitement algorithmique que soient communiquées à l'intéressé qui en fait la demande « les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre ». L'article 311-1-2 du même code exige aussi la communication d'informations encore plus précises :

> l'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :
>
> 1° Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
>
> 2° Les données traitées et leurs sources ;
>
> 3° Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;

---

[568] *Ibid*. En particulier, l'article 22 prévoit que la décision soit exclusivement basée sur un traitement automatisé, ce qui exclurait tous les processus qui font appel à un minimum d'intervention humaine.

[569] *Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (1)*, JO, 8 octobre 2016.

[570] *Code des relations entre le public et l'administration*. La loi modifie l'article 311-1-3 du Code.

4° Les opérations effectuées par le traitement[571].

L'article 21 de la *Loi relative à la protection des données personnelles*[572] en France prévoit, quant à lui que :

> le responsable de traitement s'assure de la *maîtrise* du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard [nos italiques][573].

Selon le paragraphe 71 d'une décision du Conseil Constitutionnel de France[574], l'article 21 susmentionné implique que le traitement requiert un contrôle et une validation par l'humain, ce que Beaudouin et ses coauteurs interprètent comme une prohibition de certaines techniques d'apprentissage automatique en l'absence d'outils d'explication, en l'occurrence les systèmes autonomes dont le modèle s'adapte en cours de production[575] :

> Il en résulte que ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de

---

[571] *Ibid*.

[572] *LPRP*, *supra* note 458.

[573] *Ibid*, art 21.

[574] Cons const, 12 juin 2018, [2018], 2018-765 DC.

[575] Beaudouin et al, « Flexible and Context-Specific AI Explainability », *supra* note 308 à la p 31 : « The use of convolutional neural networks would appear prohibited in the absence of robust explanation tools ».

> réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement[576].

9.3.3 Proposition de règlement du Parlement européen et Résolution du Parlement européen sur l'utilisation de l'intelligence artificielle en droit pénal

Les travaux de la Commission européenne précisent plusieurs exigences concernant l'utilisation de l'intelligence artificielle et encadrent des technologies spécifiques telles que la reconnaissance faciale, dans des contextes spécifiques tels que le droit pénal. Il n'est pas impensable que ces travaux mènent à consolider le rôle de l'Europe en tant que précurseur de la réglementation sur l'intelligence artificielle.

La *Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union* établit un cadre réglementaire minimal pour l'ensemble de l'Europe selon une approche fondée sur le risque [577] . Les exigences pour les systèmes considérés à haut risque incluent notamment :

- La mise en place d'un processus itératif et continu de gestion des risques, incluant les mesures pour atténuer ou éliminer le risque [578];
- les tests doivent être effectués sur la base de métriques et seuils probabilistes préalablement définis selon le niveau de risque du système[579];

---

[576] Cons const, 12 juin 2018, [2018], 2018-765 DC, *supra* note 558 au para 71.

[577] CE, Commission, *supra* note 549 à la p 3.

[578] CE, Commission, *supra* note 549 art 9.

[579] *Ibid* art 9.

- les jeux de données d'entraînement, validation et tests doivent être pertinents, représentatifs, exempts d'erreurs et complets[580];
- l'utilisateur doit pouvoir interpréter les résultats et utiliser le système de manière appropriée et pour ce plusieurs informations sont requises telles que les capacités, limites et performance du système pour les personnes ou groupes de personnes auxquels le système est destiné[581];
- les décisions concernant certains systèmes d'identification biométriques prises par un utilisateur sur la base du résultat du système doivent être validées par au moins deux personnes physiques [582];
- documentation technique détaillée des éléments du système et du processus de développement, incluant les choix de conception, hypothèses, compromis techniques, ce que le système est conçu pour optimiser, les choix de classification, les procédures d'étiquetage et nettoyage de données, l'évaluation des mesures de contrôle par l'humain, les paramètres utilisés pour évaluer l'exactitude, les capacités et limites du système en termes de performance y compris le niveau global d'exactitude et le niveau d'exactitude pour des groupes de personnes spécifiques[583];
- les systèmes autonomes (qui continuent d'apprendre après la mise en service) doivent faire l'objet de mesures traitant les boucles de rétroaction qui amplifient la discrimination[584];
- des mesures doivent être en place pour maîtriser les attaques malveillantes et résister aux défaillances ou incohérences provenant du système lui-même ou son environnement [585];

---

[580] *Ibid* art 10.

[581] *Ibid* art 13,14.

[582] *Ibid* art 14.

[583] *Ibid* art 11 et annexe iv.

[584] *Ibid* art 15 par 5.

[585] *Ibid* art 15.

- les systèmes doivent être soumis à une évaluation de la conformité[586] un marquage CE doit être apposé afin d'indiquer la conformité au présent règlement [587].

Le Parlement européen s'est penché sur l'utilisation de l'intelligence artificielle dans divers domaines, dont le droit pénal[588]. Ces travaux ont mené à la *Résolution du Parlement européen adoptée le 6 octobre 2021 portant sur l'utilisation de l'IA par les autorités policières et judiciaires dans les affaires pénales* [589] . La résolution recommande notamment que :

- les autorités policières et judiciaires respectent des normes juridiques extrêmement élevées étant donné le niveau de risque élevé que pose l'utilisation de ces systèmes sur les personnes concernées et le potentiel discriminatoire des systèmes[590];
- les systèmes respectent les principes d'explication [591];
- les décisions qui produisent des effets juridiques ou similaires soient toujours prises par un être humain[592];

---

[586] *Ibid* art 19.

[587] *Ibid* art 16.

[588] *Ibid* aux pp 2-3.

[589] *Résolution sur l'IA en droit pénal*, *supra* note 186.

[590] *Ibid* au para O, 8, 16.

[591] *Ibid* au para 4.

[592] *Ibid* au para 16.

- le personnel suive obligatoirement une formation spécialisée *considérable* sur les limites, risques, y compris les risques de biais et la bonne utilisation des systèmes[593];
- les systèmes soient accompagnés de documentation obligatoire exposant leurs fonctions, capacités, limites, éléments déterminants d'une décision[594], risques, performances et conditions de fonctionnement prévues[595];
- les données d'entraînements soient conservées et documentées, incluant leur contexte, finalité, exactitude, conséquences indirectes, traitement de la part des concepteurs et conformité aux droits fondamentaux [596];
- une partie indépendante réalise des audits périodiques et obligatoires afin de tester, évaluer les systèmes et veiller à ce qu'ils fonctionnent comme prévu sans effets indésirables ou préjudiciables[597];
- les systèmes d'intelligence artificielle soient accessibles, contrôlables et utilisent des logiciels libres dans la mesure du possible[598].

La résolution prend position sur diverses applications :

- l'interdiction d'applications de reconnaissance d'affects car leur validité scientifique n'a pas été reconnue [599];

---

[593] *Ibid* au para 23.

[594] *Ibid* au para 19.

[595] *Ibid* au para 17.

[596] *Ibid* au para 19.

[597] *Ibid* au para 18.

[598] *Ibid* au para 17.

[599] *Ibid* au para 30.

- le moratoire dans le déploiement des systèmes de reconnaissance faciale à cause, entre autres, du potentiel d'obtenir des résultats biaisés [600];
- l'opposition à l'utilisation de systèmes de prédiction basés sur des données historiques ou des comportements passés, ceux-ci n'étant ni fiables ni efficaces[601].

9.4 États-Unis

9.4.1 The Algorithmic Accountability Act

The Federal Rules of Evidence[602] est l'instrument qui régit la preuve en cour fédérale aux États-Unis et en particulier, la règle 702[603] s'applique aux témoignages experts [604]. Ainsi, pour être admissible, le témoignage doit être basé sur des principes et des méthodes *fiables*, ce que l'on pourra évaluer à la lumière des critères *Daubert* [605] que nous avons identifiés au Chapitre 1 de ce mémoire[606]. L'apprentissage automatique

---

[600] *Ibid* au para 27.

[601] *Ibid* au para 24.

[602] *Federal Rules of Evidence*, *supra* note 80.

[603] *Ibid* : « Rule 702 - Testimony by Expert Witnesses ».

[604] Nutter, « Machine Learning Evidence », *supra* note 117 à la p 931.

[605] *Daubert*, *supra* note 78.

[606] On rappellera que ces critères sont l'acceptation générale, la revue par les pairs, les tests et l'incertitude.

en soi est admissible en cour selon *The Rules of Evidence* et les critères *Daubert*, d'après Nutter[607].

Au niveau de la transparence, plusieurs lois adoptées avant la prolifération des techniques d'apprentissage automatique exigent une certaine transparence qui s'appliquera, peu importe la technologie utilisée. Par exemple, au sein de l'administration publique le *Privacy Act*[608] exige que soient divulgués le stockage des données personnelles ainsi que leur utilisation. Le *Freedom of Information Act*[609] donne accès à l'information détenue par l'administration ainsi qu'aux règles de décisions. Le *E-Government Act* [610] exige que les choix technologiques soient divulgués[611]. Dans le domaine commercial, le *Equal Credit Opportunity Act*[612] exige que soient divulguées les raisons pour lesquelles un individu se verrait refuser un prêt bancaire[613]. Selon Doshi-Velez, les techniques d'explication actuelles de l'intelligence artificielle répondent aux exigences des lois américaines dans la mesure où elles permettent d'identifier les principaux facteurs contribuant à une décision spécifique et

---

[607] Nutter, « Machine Learning Evidence », *supra* note 117.

[608] *Privacy Act*, 5 USC § 552a (1974).

[609] *Freedom of Information Act of 1966,* Pub L No 89-487, 80 Stat 250.

[610] *E-Government Act of 2002,* Pub L No 107-347, 116 Stat 2899.

[611] Kluttz, Kohli, Nitin, et Mulligan, Deirdre K., *supra* note 307.

[612] *Equal Credit Opportunity Act,* 15 USC 1691 (2011).

[613] Kluttz, Kohli, Nitin, et Mulligan, Deirdre K., *supra* note 307.

permettent des explications contrefactuelles[614]. Ces techniques permettent de justifier adéquatement une décision même si de larges pans du système restent opaques[615].

Aux lois mentionnées plus haut, s'ajoute le projet de loi *Algorithmic Accountability Act*[616] de 2019 présenté au Congrès américain et visant les compagnies qui répondent aux conditions suivantes : « 50 million in average annual gross receipts, that hold personal information of at least 1 million individuals or their devices, or that act primarily as data brokers »[617]. Ce projet de loi vise à réduire les risques de discrimination que pourrait entraîner l'usage des algorithmes d'apprentissage automatique[618]. Pour ce faire, de nouveaux droits sont accordés à la *Federal Trade Commission* des États-Unis lui donnant le pouvoir de contraindre les entreprises à

---

[614] Doshi-Velez et al, « Accountability of AI Under the Law », *supra* note 288 à la p 7.

[615] *Ibid* aux pp 6-7. Au même titre qu'il n'est pas raisonnable d'exiger de l'humain qu'il justifie une décision en relatant l'ensemble des signaux neuronaux qui y ont contribuée, il ne serait pas raisonnable non plus, selon les auteurs, d'exiger de la machine l'ensemble des flux de signaux numériques.

[616] É-U, Bill HR 2231, *Algorithmic Accountability Act*, 116e Cong, 2019.

[617] Scherman, Michael et al, « US Lawmakers propose Algorithmic Accountability Act intended to regulate AI, April 22, 2019, dans CyberLex : insights on cybersecurity, privacy and data protection law », (avril 2019), en ligne: <https://edoctrine.caij.qc.ca/publications-cabinets/mccarthy/2019/a98358/en/PC-a115788>, McCarthy Tétrault.

[618] Scherman, Michael et al, « US Lawmakers propose Algorithmic Accountability Act intended to regulate AI, April 22, 2019, dans CyberLex : insights on cybersecurity, privacy and data protection law, McCarthy Tétrault », (avril 2019), en ligne: <https://edoctrine.caij.qc.ca/publications-cabinets/mccarthy/2019/a98358/en/PC-a115788>, McCarthy Tétrault.

évaluer les risques liés à la discrimination[619] et auditer leurs systèmes au regard de la discrimination[620].

9.5 Au Canada

9.5.1 Lois sur la preuve au Canada et au Québec

Outre la common law qui s'applique à titre supplétif au Québec dans les matières d'ordre criminel[621], la *Loi sur la preuve au Canada*[622] est l'une des principales sources de droit en matière de preuve en droit criminel[623]. Or celle-ci ne fait aucune mention de l'intelligence artificielle ou systèmes automatisés et, puisque son article 40[624] précise que les lois de la province dans laquelle les procédures sont exercées s'appliquent, par exemple les procédures d'admissibilité de la preuve, il convient de se pencher, d'une part, sur le *Code civil du Québec*[625] et en particulier, son article 2843

---

[619] Éric Lavallée, « La définition juridique de l'intelligence artificielle évolue : différents pays, différentes approches », (mars 2020), en ligne: *Lavery* <https://edoctrine.caij.qc.ca/publications-cabinets/lavery/2020/a121811/fr/i73b2616b-c0b5-4351-954c-4cce21486918>.

[620] Scherman, Michael et al, *supra* note 619.

[621] Léo Ducharme, *Précis de la preuve*, 6e éd, Montréal, Wilson & Lafleur, 2005 au no 37; *Commission scolaire de Victoriaville c La Reine*, [2002] CanLII 61082 au para 44 (CCI).

[622] *Loi sur la preuve au Canada*, *supra* note 39.

[623] Bellemare, *supra* note 39 à la p 123.

[624] *Loi sur la preuve au Canada*, *supra* note 39, art 40 : « Dans toutes les procédures qui relèvent de l'autorité législative du Parlement du Canada, les lois sur la preuve qui sont en vigueur dans la province où ces procédures sont exercées, [...] s'appliquent à ces procédures ».

[625] *CcQ*, *supra* note 39. Les articles 2803 à 2874 concernent la preuve.

lequel concerne le témoignage expert[626]. Là encore, aucune mention n'y est faite de l'intelligence artificielle. D'autre part, toujours au Québec, la *Loi concernant le cadre juridique des technologies de l'information*[627] qui traite de la preuve documentaire précise notamment la valeur juridique de documents sur support numérique (par opposition au support papier) ainsi que les moyens permettant de les relier à une personne[628] ou d'identifier une personne[629]. Notons que la loi précise à l'article 63 qu'un comité multidisciplinaire soit constitué « pour favoriser l'harmonisation […] des procédés, systèmes, normes et standards techniques » reliés à ces objectifs[630]. On pourrait s'inspirer d'un tel comité pour se pencher sur la fiabilité des résultats produits par les systèmes d'apprentissage automatique en preuve qui, dans un contexte de droit criminel dont le Code s'applique à l'ensemble du Canada, assurait d'une harmonisation pancanadienne, ce qui éviterait les disparités au pays[631].

---

[626] *Ibid*, art 2843 : « Le témoignage est la déclaration par laquelle une personne relate les faits dont elle a eu personnellement connaissance ou par laquelle un expert donne son avis. Il doit, pour faire preuve, être contenu dans une déposition faite à l'instance, sauf du consentement des parties ou dans les cas prévus par la loi ».

[627] *Loi concernant le cadre juridique des technologies de l'information*, RLRQ, c C-1.1.

[628] *Ibid*, arts 38-39.

[629] *Ibid*, arts 41-46.

[630] *Ibid*, arts 63-68.

[631] La Conférence pour l'harmonisation des lois au Canada pourrait se révéler un bon véhicule pour aborder cet enjeu: « Conférence pour l'harmonisation des lois au Canada », en ligne: <https://www.ulcc-chlc.ca/?lang=fr-ca>.

9.5.2 La Directive sur la prise de décision automatisée au Canada

La *Directive sur la prise de décision automatisée*[632] publiée par le Conseil du trésor du Canada et entrée en vigueur en 2019 concerne l'ensemble de l'appareil administratif canadien. Ses exigences seront plus contraignantes selon l'augmentation du niveau d'incidence du système sur les personnes, les collectivités et les écosystèmes[633].

La Directive couvre plusieurs facteurs de fiabilité discutés dans notre ouvrage, notamment :

- La pertinence, exactitude, mise à jour et conformité des données[634]
- Une explication s*ignificative* sur la façon dont la décision a été prise et la raison pour laquelle elle a été prise[635]
- Une formation aux membres du personnel leur permettant d'examiner, expliquer et surveiller le système[636]
- L'absence de biais[637]

---

[632] Canada, Conseil du Trésor, *Directive sur la prise de décision automatisée*, 2019.

[633] *Directive sur la prise de décision automatisée*, 2019. Les exigences sont déterminées selon quatre niveaux d'incidence. Voir les annexes B et C. Canada, Secrétariat du Conseil du Trésor du, « Outil d'évaluation de l'incidence algorithmique », (22 mars 2021), en ligne: <https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html>.

[634] Canada, Conseil du Trésor, *supra* note 525 à l'exigence 6.3.3.

[635] *Ibid* à l'exigence 6.2.3.

[636] *Ibid* à l'exigence 6.3.4.

[637] *Ibid* à l'exigence 6.3.1.

- L'intervention humaine obligatoire pour les systèmes à incidence élevée, dont la prise de décision finale[638]
- Un processus d'assurance qualité avant et durant la production[639]
- La revue par les pairs et l'accès aux composantes logiciels, incluant le code source[640]
- Un plan de gestion de risques[641]

9.5.3 La Loi sur la protection des renseignements personnels

Au Canada les lois sur la protection des renseignements personnels ne sont pas spécifiques à l'intelligence artificielle mais s'appliquent néanmoins, peu importe le type de technologies utilisées[642]. La *Loi sur la protection des renseignements personnels*[643] s'applique aux services gouvernementaux fédéraux, tels que les services de police fédérale et de sécurité publique et la sécurité frontalière[644] tandis que la *Loi*

---

[638] *Ibid* aux exigences 6.3.9 et 6.3.10.

[639] *Ibid* à l'exigence 6.3.2.

[640] *Ibid* aux exigences 6.2.4. à 6.2.6 et 6.3.4.

[641] *Ibid* à l'exigence 6.3.7.

[642] Guillaume Laberge et Éric Lavallée, « L'intelligence artificielle, bientôt réglementée au Canada ? », (29 janvier 2021), en ligne: *Lavery* <https://edoctrine.caij.qc.ca/publications-cabinets/lavery/2021/a121811/fr/i3fbc498f-432f-46da-8675-a5e7ec521452>. On y retrouve la Loi sur la protection des renseignements personnels et la Loi sur la protection des renseignements personnels et des documents électroniques. *Loi sur la protection des renseignements personnels*, LRC 1985, ch. P-21; *La Loi sur la protection des renseignements personnels et les documents électroniques*, LC 2000, c 5.

[643] *LPRP*, *supra* note 458.

[644] Canada, Commissariat à la protection de la vie privée, « Survol de la Loi sur la protection des renseignements personnels », (23 août 2019), en ligne: <https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/la-loi-sur-la-protection-des-renseignements-personnels/pa_brief/>.

sur la protection des renseignements personnels et des données électroniques (LPRPDE)[645] s'applique à toutes les « organisations qui recueillent, utilisent ou communiquent des renseignements personnels dans le cadre de leurs activités commerciales », incluant « les organisations sous réglementation fédérale qui exercent leurs activités au Canada » telles que les entreprises de télécommunication, les aéroports, lignes aériennes et banques[646].

La LPRPDE a fait l'objet de propositions stratégiques pour une réforme en 2020 dans laquelle les enjeux de l'intelligence artificielle sont explicitement abordés[647]. En ce qui a trait à la fiabilité, mentionnons en particulier :

- La reconnaissance du potentiel discriminatoire des systèmes de décisions automatisés[648];
- Le recours à une explication *valable*[649], laquelle « permet aux individus de comprendre la nature et les éléments de la décision dont ils font l'objet ou les

---

[645] *LPRPDE*, *supra* note 302.

[646] Canada, Commissariat à la protection de la vie privée, « Survol de la LPRPDE », (1 mai 2019), en ligne: <https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/la-loi-sur-la-protection-des-renseignements-personnels-et-les-documents-electroniques-lprpde/lprpde_survol/>.

[647] Guillaume Laberge et Éric Lavallée, *supra* note 643.

[648] Cofone, *supra* note 302 au para 4a.

[649] Le professeur Cofone, auteur des recommandations pour la réforme, afin de contrer les risques de discrimination, recommande « [le] droit d'obtenir une explication valable lorsqu'un individu fait l'objet d'une prise de décision automatisée qui utilise des renseignements personnels le concernant ». Voir Cofone, *supra* note 302 à la recommandation 11 de l'annexe.

règles qui définissent le traitement et les caractéristiques principales de la décision »[650];

- Le droit de corriger les renseignements personnels inexacts, y compris les inférences[651].

De plus, reconnaissant que l'explication pourra dépendre du système mais cherchant néanmoins à conserver la neutralité technologique de la LPRPDE, il est recommandé de « publier une ligne directrice ou un règlement qui clarifie le sens et l'ampleur de l'explication requise de manière à préciser les explications qui sont pertinentes pour les différentes technologies »[652].

Ces recommandations s'appliqueraient sans distinction sur le degré d'intervention humaine dans le processus de décision[653]. En effet, parlant du terme *prise de décision automatisée*, le professeur Cofone précise : « des mots tels qu'« uniquement » et « exclusivement » ne devraient pas être juxtaposés à ce terme, puisque cela aurait pour effet de restreindre considérablement l'applicabilité de certaines mesures de protection »[654].

---

[650] *Ibid* au para 4b.

[651] *Ibid* au para 2c.

[652] *Ibid* au para 4b.

[653] Cofone, *supra* note 302. Cette recommandation se distingue de la *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels* au Québec et du RGPD (Règlement général sur la protection des données) en Europe.

[654] Canada, Commissariat à la protection de la vie privée, « Un cadre réglementaire pour l'IA : recommandations pour la réforme de la LPRPDE », (12 novembre 2020), en ligne: <https://www.priv.gc.ca/fr/a-propos-du-commissariat/ce-que-nous-faisons/consultations/consultations-terminees/consultation-ai/reg-fw_202011/#fn10>.

9.5.4 La Loi modernisant des dispositions législatives en matière de protection des renseignements personnels du Québec

Au Québec, la *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels* [655] concerne les traitements automatisés utilisés par les organismes publics et les entreprises[656]. La loi prévoit :

- L'explication d'une décision[657],
- Le droit de corriger les renseignements personnels[658].

En effet, l'article 21 spécifie qu'un organisme public qui fournit « une décision fondée *exclusivement* sur un traitement automatisé » [nos italiques][659] doit informer la personne concernée : « 1° des renseignements personnels utilisés pour rendre la décision; 2° des raisons, ainsi que des principaux facteurs et paramètres, ayant mené à la décision; 3° de son droit de faire rectifier les renseignements personnels utilisés pour rendre la décision »[660]. L'article 110 pose les mêmes exigences aux entreprises privées[661]. La loi n'offre pas de détails quant à l'application de ses dispositions, par

---

[655] *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*, LQ 2021, c 25, 2021.

[656] *Ibid* aux pp 2-3. Voir les notes explicatives.

[657] *Ibid*, art 21.

[658] *Ibid*, art 20.

[659] *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*, LQ 2021, c 25, *supra* note 656.

[660] *Ibid*, art 21.

[661] *Ibid*, art 110. Voir à la page 40.

exemple elle est silencieuse au sujet du type de renseignements personnels ou principaux facteurs menant à une décision[662]. Elle ne précise pas plus si l'explication doit porter sur le fonctionnement général du système ou la décision spécifique[663].

Tel que mentionné plus haut, la portée de la loi est considérablement restreinte par le mot « exclusivement », selon le Professeur Cofone. On pourrait dès lors considérer qu'une intervention humaine dans le processus de décision exonère des exigences ci-haut.

---

[662] Henri Vanessa, Deneault-Rouillard William et Agaby Linda, « Projet de loi n° 64 : Nouvelles règles encadrant la prise de décision individuelle automatisée. », (octobre 2020), en ligne: *Fasken* <https://edoctrine.caij.qc.ca/publications-cabinets/fasken/2020/a121765/fr/iac9cf9f8-7d17-47b1-9ff4-c61a33ca0af0>. Bien que l'article de Henri et ses coauteurs se penche sur le projet de loi 64, la version adoptée en Septembre 2021 ne comprend pas d'amendements qui touchent les dispositions dont il est question. Voir Stoddart, Jennifer, Uzan-Naulin, Julie, et Romano, Mathilde, *Le début d'un temps nouveau pour le secteur privé : le projet de loi 64 sur la protection des renseignements personnels est adopté*, Fasken, 2021. Voir le tableau des amendements à la fin de l'article.

[663] Vanessa, William et Linda, *supra* note 663. Stoddart, Jennifer, Uzan-Naulin, Julie, et Romano, Mathilde, *supra* note 663.

CHAPITRE X

EXEMPLE D'ÉVALUATION DE LA FIABILITÉ D'UN OUTIL DE RECONNAISSANCE DU LOCUTEUR

10.1 La reconnaissance du locuteur

Dans ce chapitre, nous présentons un exemple fictif de l'évaluation de la fiabilité d'un outil (fictif) de reconnaissance du locuteur. Distinguons tout d'abord la *reconnaissance du locuteur* de la *reconnaissance de la parole*. Cette dernière vise à déchiffrer le contenu et non l'identité du locuteur. Des exemples d'applications de reconnaissance de la parole sont les assistants vocaux tels que Siri[664], Alexa[665] ou

[664] Apple, « Siri », (mai 2021), en ligne: *Apple (CA)* <https://www.apple.com/ca/fr/siri/>.

[665] Amazon, « Amazon.ca Aide: FAQ sur Alexa et les appareils Alexa », (novembre 2021), en ligne: *Amazon* <https://www.amazon.ca/-/fr/gp/help/customer/display.html?nodeId=201602230>.

l'Assistant Google[666] ou des applications telles que Dragon[667] qui convertissent la voix en texte.

Lorsque l'on vise à confirmer qu'une personne est bien celle qu'elle prétend être, on parlera de *vérification vocale*[668]. On cherche donc à authentifier une personne en comparant deux échantillons de voix. La vérification vocale peut être utilisée comme procédure de sécurité pour remplacer le mot de passe, par exemple, lors d'un accès à un compte bancaire par téléphone[669]. *L'identification du locuteur* vise à identifier un locuteur inconnu parmi plusieurs locuteurs connus. On compare alors un échantillon avec plusieurs échantillons de locuteurs connus[670]. Dans les deux cas, on pourra parler de *reconnaissance du locuteur*[671].

Supposons qu'un enregistrement téléphonique dans lequel on entend plusieurs voix soit soumis pour analyse à un laboratoire de police scientifique afin d'identifier le locuteur qui y aurait prononcé une phrase incriminante. Pour identifier le locuteur, la

---

[666] Google, « Assistant Google – Votre Google personnel », (novembre 2021), en ligne: *Google* <https://assistant.google.com/intl/fr_ca/>.

[667] Nuance Communications, « Dragon Speech Recognition - Get More Done by Voice | Nuance », (2021), en ligne: *Nuance Communications* <https://www.nuance.com/dragon.html>.

[668] Ajili, *supra* note 236 à la p 70.

[669] L'exemple est tiré du site de Nuance qui offre des solutions de vérification vocale : Nuance Communications, « Biometric Authentication | Strong Customer Authentication », (2021), en ligne: *Nuance Communications* <https://www.nuance.com/omni-channel-customer-engagement/authentication-and-fraud-prevention/biometric-authentication.html>.

[670] Ajili, *supra* note 236 à la p 70.

[671] *Ibid* à la p 71.

police scientifique pourrait effectuer un prélèvement sonore directement sur le suspect[672], obtenir un enregistrement alors qu'il est détenu[673] ou d'une mise sur écoute téléphonique du suspect[674], ou la police pourra utiliser une banque de données recueillies à partir de réseaux sociaux, par exemple[675]. Le premier échantillon contient les paroles incriminantes. Le second échantillon contient la voix du suspect. La police comparera alors les deux enregistrements à l'aide d'un outil d'apprentissage automatique. Deux outils courants utilisés en reconnaissance du locuteur, selon un sondage sur les pratiques de la police scientifique[676], sont Batvox[677] et Nuance

---

[672] Ce cas de figure est relaté par Christophe Steccoli de la Police scientifique dans France Inter, « Épisode 6: Chacun sa voix du 01 août 2015 », (août 2015), en ligne: *France Inter* <https://www.franceinter.fr/emissions/scenes-de-crime/scenes-de-crime-01-aout-2015>.

[673] Ce cas de figure est évoqué dans cet article : Alice Moreno, « Police scientifique : comment les voix et les sons sont analysés pour résoudre des enquêtes », (28 octobre 2020), en ligne: *RTL* <https://www.rtl.fr/actu/justice-faits-divers/police-scientifique-comment-les-voix-et-les-sons-sont-analyses-pour-resoudre-des-enquetes-7800912370>.

[674] Ce cas de figure est tiré de l'affaire Bardon-Kulik qui a eu un fort retentissement en France alors que l'identification vocale était en cause : Ouest-France avec AFP, « "On l'entend mourir" : le terrible appel aux secours d'Élodie Kulik diffusé au procès de Willy Bardon », (27 novembre 2019), en ligne: <https://www.ouest-france.fr/societe/justice/l-entend-mourir-le-terrible-appel-aux-secours-d-elodie-kulik-diffuse-au-proces-de-willy-bardon-6627833>.

[675] Khelif et al, *supra* note 367. Le projet européen SIIP (Speaker Identification Integrated Project), destiné aux forces de l'ordre, vise la collecte de données de sources variées pour constituer une large banque de voix.

[676] Erica Gold et Peter French, « International Practices in Forensic Speaker Comparisons: Second Survey » (2019) 26:1 International Journal of Speech, Language and the Law 1.

[677] Scientific Analytical Tools, « BATVOX », (2017), en ligne: *Scientific Analytical Tools* <https://sat.ae/audio-forensics/forensic-analysis-tools/batvox/>.

Forensics[678]. Le résultat de l'analyse pourra alors être présenté en preuve à la cour, incriminant ou non le suspect.

10.2 Les témoignages auditifs

Dans *R v Pinch*, la Cour supérieure de l'Ontario distingue explicitement les caractéristiques du témoignage auditif profane de celles du témoignage expert[679]. Le témoin profane est une personne familière avec la voix de la personne en question, un membre des forces de l'ordre dont l'oreille est exercée ou un juge des faits qui écoute un enregistrement admis en preuve[680]. Le témoin expert, en revanche, possède des qualifications précises, par exemple à titre de criminaliste spécialiste en phonétique, acoustique ou en linguistique[681].

Bien que notre exemple concerne le témoignage expert, nous ferons un détour par le témoignage profane pour illustrer certaines difficultés perçues par le tribunal sur témoignage auditif en général.

---

[678] Nuance Communications, « Nuance Forensics Datasheet », (2018), en ligne: *Nuance Communications* <https://www.nuance.com/content/dam/nuance/en_us/collateral/enterprise/data-sheet/ds-nuance-forensics-en-us.pdf>.

[679] *R v Pinch*, [2011] 2011 ONSC 5484 au para 78.

[680] *Ibid* au para 70.

[681] *Ibid* au para 69.

10.2.1 La fiabilité du témoignage profane

Les tribunaux canadiens ont en général accueilli le témoignage profane avec extrême prudence, acquiesçant à la faillibilité de ce type de preuve[682]. Plusieurs études empiriques démontrent en effet l'inexactitude du témoignage auditif profane qui a d'ailleurs donné lieu à plusieurs erreurs judiciaires, en particulier aux États-Unis[683]. Ainsi certains experts qualifient cette preuve de « propice à l'erreur » ou « hautement douteuse »[684]. Des études démontrent un taux de faux positifs aussi variable que 10% à 90% selon les conditions de tests[685]. Parmi les conditions qui affectent la fiabilité de la preuve, on peut noter la durée de l'exposition à la voix, l'interval de rétention (c'est-à-dire la durée entre la première exposition lors du crime et la seconde lors de l'identification), l'attention avec laquelle le témoin écoute la voix lors du crime, le médium de communication (par exemple, le téléphone) et le degré avec lequel le criminel déguise sa voix lors du crime[686]. De plus, des caractéristiques telles que la familiarité de la voix, l'accent, la langue ou le caractère distinctif de la voix (par

---

[682] Christopher Sherrin, « Earwitness Evidence: The Reliability of Voice Identifications » (2016) 52:3 Osgoode Hall LJ 819. L'auteur mentionne certains cas où la circonspection de la cour sur la fiabilité du témoignage auditif n'a pas toujours penché vers la prudence, dénotant une certaine hétérogénéité dans le traitement de ce type de preuve.

[683] *Ibid* à la p 822. L'auteur mentionne au moins deux cas au Canada et dénombre pas moins de dix-sept cas aux États-Unis.

[684] Sherrin, « Earwitness Evidence », *supra* note 683.

[685] *Ibid* à la p 826.

[686] Sherrin, « Earwitness Evidence », *supra* note 683.

exemple, une voix nasillarde), ne sont pas des gages de fiabilité du témoignage ainsi que le démontrent maintes études[687].

La reconnaissance d'une voix par un profane, même familier avec la voix, pose donc des enjeux de fiabilité reconnus par les tribunaux canadiens.

10.2.2 Les outils du témoignage expert

Parmi les outils utilisés par le témoin expert, le « spectrographe acoustique » a couramment été soumis en preuve au tribunal[688]. Le spectrographe est une technique mise au point lors de la deuxième guerre mondiale et qui avait comme objectif l'identification des ennemis à partir d'ondes radio[689]. Le spectrographe est une représentation graphique des variations de fréquences émises par les sons produits par la voix sur une échelle de temps[690].

Aux États-Unis, l'analyse spectrographique fut admise dans plusieurs cours dans les années 1960 et 1970 jusqu'à ce qu'une étude de la National Academy of Science en 1979 remette en question la fiabilité de cette technique[691]. Cette étude mit un frein

---

[687] *Ibid*.

[688] Daniel A Bellemare, « L'identification d'un accusé par ses vocogrammes » (1978) 56:3 R du B can 440.

[689] Anne Karpf, *La voix. Un univers invisible*, Hors collection, Paris, Autrement, 2008 aux pp 427-457. Voir le chapitre « 16 - Empreintes vocales et vol de la voix » .

[690] Bellemare, *supra* note 689.

[691] Lawrence Solan et PM Tiersma, « Hearing voices: Speaker identification in court » (2003) 54:2 Hastings LJ 373 à la p 376.

à son admissibilité devant les tribunaux, à quelques exceptions près[692]. Au Canada, le professeur Patenaude se demandera en 2002 s'il était bien « sage » pour la Cour d'admettre en preuve le spectrographe dans des affaires comme *R. c. Montani*[693] et *R. c. Medvedew*[694], deux affaires qui se déroulèrent avant la publication de l'étude de 1979[695]. L'IAFPA (International Association for Forensics Phonetics and Acoustics) dans une résolution de 2007 a formellement rejeté la comparaison visuelle de représentation spectrale considérant qu'elle était sans fondement scientifique et ne devrait pas être utilisée pour des analyses de cas forensiques[696].

Les outils plus récents de reconnaissance du locuteur utilisent des techniques d'apprentissage automatique. L'outil Nuance Forensics, par exemple, est basé sur l'apprentissage profond[697]. Selon la compagnie Nuance Communications, l'outil serait utilisé dans plus de 40 pays[698], dont la Gendarmerie royale du Canada (GRC)[699]. Ce

---

[692] *Ibid*.

[693] Tel que cité dans Patenaude, *supra* note 79 à la p 113. Voir la note 5 : « (1974) 26 C.R. (N.S.) 339 (Ont. Prov. Ct.) ».

[694] *R v Medvedew*, 91 DLR (3d) 21, 6 CR (3d) 185, 1978 CanLII 2101 n 6.

[695] Patenaude, *supra* note 79.

[696] Louis-Jean Boë et Jean-François Bonastre, « L'identification du locuteur : 20 ans de témoignage dans les cours de Justice. Le cas du Lipsadon » (2012) Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP 417-424, art 2.3.8.

[697] Nuance Communications, *supra* note 679.

[698] *Ibid*.

[699] Almog Aley-Raz, *«Voice Biometrics, The Silent Revolution, Commercial and Government success stories»,* Israel's Biometrics and Strong Authentification Conference, 2016 à la p 24.

genre d'outil a été soumis en preuve en cour fédérale aux États-Unis pour la première fois en 2015 lors d'une affaire concernant le groupe militant al-Shabaab en Somalie selon le Wall Street Journal[700]. En soumettant la preuve, la poursuite notait devant la cour fédérale que ce genre d'outil avait été admis lors de procédures judiciaires auparavant devant une cour d'État aux États-Unis et devant plusieurs cours européennes[701]. Dans cette affaire, deux autres expertises avaient été demandées, soit une analyse acoustique qui mesure la vitesse à laquelle le locuteur parler et une analyse phonétique qui compare la prononciation de certains sons[702].

En France, les scientifiques spécialistes de la parole au sein de la Société française d'acoustique ont émis une motion en 1990 affirmant que l'identification d'un individu par la voix était un problème non résolu[703]. En 1997, la motion fut renforcée par une pétition demandant un moratoire sur les expertises produites pour la justice tant que la technique n'était pas validée scientifiquement. La pétition a eu des répercussions au sein des tribunaux. Ainsi les chercheurs académiques ne se présentent plus comme *experts judiciaires en reconnaissance de la voix* lorsqu'ils témoignent devant les tribunaux, mais interviennent plutôt à titre de *témoin scientifique*, ce qui permet

---

[700] Hong, *supra* note 16.

[701] *Ibid*.

[702] *Ibid*.

[703] Jean-François Bonastre, « 1990-2020 : retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire » dans Gilles Adda, Maxime Amblard et Karën Fort, dir, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, Nancy, France, ATALA, 2020, 38.

d'expliquer aux tribunaux l'état actuel de la science en ce qui concerne la reconnaissance vocale[704]. La motion est toujours en vigueur en 2022[705].

10.3 Les enjeux de la reconnaissance par la voix

Dans les années 1960, le terme « empreinte vocale » a été accolé au spectrographe, par association aux « empreintes digitales ». Cependant, le terme « empreinte » appliqué à la voix est abusif car l'identification par empreinte vocale est loin d'être aussi exacte que l'identification par empreinte digitale selon plusieurs experts[706]. En effet, une personne ne prononce jamais le même mot deux fois de la même façon, y compris dans la même phrase[707]. La parole relève de la dynamique de la « performance », laquelle variera selon le contexte, par exemple la lecture d'un texte ou la discussion spontanée[708]. La voix changera également selon des caractéristiques physiques et psychologiques telles que l'humeur, la fatigue, l'âge ou l'état de santé[709].

---

[704] *Ibid*.

[705] Selon une communication privée avec Jean-François Bonastre, auteur de *Ibid*.

[706] Bellemare, *supra* note 689 aux notes 2 et 9; *United States v Baller*, 519 F (2d) 463, 1975 US App Lexis 13781. Dans cette affaire, le juge Peck dira: « The use of the term "voiceprint", with its overtones of "fingerprint", gives voice spectrographic identification an aura of absolute certainty and accuracy which is neither justified by the facts nor claimed by experts in the field ».

[707] Karpf, *supra* note 690.

[708] Craig S Greenberg et al, « Two Decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology » (2020) Computer Speech and Language, en ligne: <https://www.nist.gov/publications/two-decades-speaker-recognition-evaluation-national-institute-standards-and-technology> à la p 2.

[709] Félicien Vallet, « Jean-François Bonastre : "La voix n'est pas une biométrie classique" », (février 2017), en ligne: *Laboratoire d'innovation numérique de la CNIL* <https://linc.cnil.fr/fr/jean-francois-bonastre-la-voix-nest-pas-une-biometrie-classique>; Reda Jourani, *Reconnaissance automatique du*

La voix est liée au langage et chargée de communiquer les données les plus diverses, des pensées aux sentiments, ajoutant aux difficultés de la modéliser[710]. On pourra même fausser sa voix en adoptant une démarche ou en parlant de façon spécifique[711] ou à l'aide d'outils automatisés[712]. À ces facteurs qui impactent la variabilité de la voix, s'ajoutent les facteurs qui complexifient la reconnaissance de la voix, soient les facteurs linguistiques tels que l'accent ou le vocabulaire, les facteurs techniques de prélèvement de la voix tels que les microphones ainsi que les conditions d'enregistrement, tels que le bruit ambiant ou la proximité du locuteur[713]. Finalement, la reconnaissance de la voix suppose que notre voix est unique bien qu'aucune démonstration n'ait permis de confirmer cette hypothèse, laquelle reste donc *plausible* tout au plus[714]. Ainsi certains diront que la fiabilité de l'analyse de la voix est plus proche de celle du détecteur de mensonges que de celle de l'empreinte digitale[715].

---

locuteur par des GMM à grande marge, thèse de doctorat en informatique, Université Paul Sabatier - Toulouse III, 2012 [non publiée]; Karpf, *supra* note 690.

[710] Félicien Vallet, *supra* note 710.

[711] *Ibid*.

[712] Voir cet article pour plus de détails sur les outils automatisés : Linlin Zheng et al, « When Automatic Voice Disguise Meets Automatic Speaker Verification » (2021) 16 IEEE Transactions on Information Forensics and Security 824.

[713] Ajili, *supra* note 236 à la p 50.

[714] *Ibid* à la p 75.

[715] La *Commission d'accès à l'information* du Québec classe l'empreinte vocale parmi la biométrie comportementale par opposition à la biométrie morphologique (par exemple, les empreintes digitales) ou biologique (par exemple, l'ADN). Voir Québec, Commission d'accès à l'information du Québec, « Biométrie », (mars 2021), en ligne: <https://www.cai.gouv.qc.ca/biometrie/>; Karpf, *supra* note 690.

Fort de ces connaissances sur les enjeux de la reconnaissance par la voix, de son historique au sein des tribunaux et des prises de position de communautés scientifiques, nous nous penchons maintenant sur un exemple de reconnaissance du locuteur par un outil d'apprentissage automatique.

10.4 Le cas

Le cas présenté est fictif, de même que les questions et réponses entre avocats et expert. Le cas est inspiré de l'affaire *Kulik-Bardon*[716] en France et de l'affaire *Slade*[717] au Royaume-Uni. Les réponses de l'expert sont en grande partie tirées d'ouvrages scientifiques[718] et du document commercial de Nuance Forensics[719]. Les questions et réponses sont présentées en bloc pour faciliter la lecture.

En 2015, un appel au 9-1-1 est enregistré. L'appel provient du téléphone cellulaire de la victime, Bobby, retrouvée avec une balle dans la tête quelques minutes plus tard par la police envoyée sur les lieux. L'enregistrement, qui constituera l'échantillon 1, contient plusieurs voix qui se superposent et beaucoup de bruits de fond où on entend ce qui semblerait être des bruits de lutte entre plusieurs hommes et d'objets cassés. On y entend la phrase « Tu l'as cherché pis là, m'a t'tuer, Bobby ». On cherche à identifier le locuteur de cette phrase. Le second échantillon a été pris

---

[716] Ouest-France avec AFP, *supra* note 675.

[717] *R v Slade & Others*, [2015] EWCA Crim 71 .

[718] Jourani, *supra* note 710; Ajili, *supra* note 236; Geoffrey Stewart Morrison, « Admissibility of forensic voice comparison testimony in England and Wales » (2018) 2018:1 Crim L Rev 20; Gold et French, « International Practices in Forensic Speaker Comparisons », *supra* note 677.

[719] Nuance Communications, *supra* note 679.

directement auprès du suspect lors d'un prélèvement sonore dans les bureaux de la police au cours du dernier mois. Le prélèvement a été réalisé par le témoin expert en vue de l'analyse vocale. On a ensuite demandé à l'expert d'analyser les deux échantillons pour savoir s'il s'agissait bien du même locuteur, en d'autres mots, si le suspect a bien prononcé la phrase « Tu l'as cherché, pis là, m'a t'tuer, Bobby » il y a cinq ans, sur la scène du crime.

10.4.1 Fiabilité du principe fondamental

10.4.1.1 Questions

- Quelle est le principe derrière un système automatisé reconnaissance vocale?
- Quel est l'objectif précis du système?
- Quelles sont les hypothèses *a priori* utilisées?
- Où intervient l'humain versus le système?
- Les procédures sont-elles standardisées?

10.4.1.2 Réponses

Le système vise à permettre d'évaluer si le locuteur sur un enregistrement est le même que le locuteur sur un autre enregistrement. Pour se faire, le système extrait les caractéristiques individuelles de la voix. En effet, l'anatomie d'un individu, la configuration précise de son conduit vocal, donne lieu à des différences acoustiques qui reflètent une structure de fréquences lorsque l'individu parle. À partir de ces caractéristiques, le système produit un modèle statistique de la voix. Les modèles sur les deux enregistrements sont ensuite comparés. Pour savoir si leurs différences et similitudes sont significatives, elles sont mesurées par rapport à une population de référence. En d'autres mots, le locuteur inconnu est-il suffisamment typique par rapport à une population de référence?

Le système produira un rapport de vraisemblance qui évalue la probabilité de deux hypothèses : soit le même locuteur se trouve sur les deux échantillons, soit ce sont

des locuteurs distincts. Aucune hypothèse *a priori* n'entre en compte dans le résultat. Seules les voix sur les enregistrements sont considérées.

L'extraction des caractéristiques de la voix, la modélisation de la voix et la comparaison entre les modèles de voix sont automatisées. L'humain intervient au début pour valider la qualité des enregistrements et les convertir dans le format supporté par le système. Il peut arriver que l'échantillon n'ait pas la qualité requise pour l'analyser, auquel cas l'analyse n'est pas effectuée. Nous convertissons ensuite l'échantillon dans le format supporté par le logiciel. Nous pouvons couper certains morceaux superflus. Cette procédure de préanalyse est documentée et le résultat est systématiquement vérifié par un collègue.

10.4.2 Fiabilité de la méthode

10.4.2.1 Questions sur les enregistrements

- Quelles sont les caractéristiques des deux enregistrements et comment se qualifie la qualité des enregistrements?
- Les enregistrements ont-ils pu se dégrader?
- Les enregistrements ont-ils été modifiés et si oui, est-ce que ces modifications ont pu impacter les caractéristiques de la voix?
- Les enregistrements étaient-ils sécurisés, à l'abri d'accès frauduleux?

10.4.2.2 Réponses

L'enregistrement 1 a une durée de moins de 10 secondes à l'origine, l'appel au 9-1-1 a été enregistré au complet. L'enregistrement a été manipulé pour le convertir dans un format accepté par le système de reconnaissance vocale et édité pour ne laisser que la phrase incriminante dont la durée est de trois secondes. La voix n'est pas altérée par ces manipulations. Le locuteur crie et semble loin du téléphone. L'échantillon 1 contient beaucoup de bruits de fond mais aucune autre voix ne semble s'y superposer.

La qualité de l'enregistrement est moyenne. Sa durée, soit trois secondes, est le minimum recommandé par le système.

Le second enregistrement a une durée de 30 minutes à l'origine. Nous l'avons édité pour ne laisser que la voix du suspect. Nous avons 20 minutes de matériel. Le micro est de calibre professionnel et a été placé à 20 cm du suspect. Il n'y a aucun bruit de fond ou voix superposée.

Nous travaillons avec des copies des enregistrements d'origine. Celles-ci sont fidèles aux originaux et non dégradées. La voix n'a pas été modifiée par nos manipulations. Les enregistrements sont numériques, localisés sur nos serveurs qui sont sécurisés d'accès frauduleux par un tiers.

10.4.2.3 Questions sur le rapport de vraisemblance

- Quelles sont les deux hypothèses du rapport de vraisemblance?
- Quelle est la valeur du rapport de vraisemblance?
- Comment traduit-on verbalement le rapport de vraisemblance en utilisant une échelle telle que celle présentée à la Figure 3?

10.4.2.4 Réponse

Le résultat produit par le système est un rapport de vraisemblance où la force de deux hypothèses est comparée étant donné les résultats d'analyse (E) :

- $Rapport\ de\ vraisemblance = \frac{P(E|H1)}{P(E|H2)}$

L'hypothèse 1 (H1) est que la voix de l'échantillon 1 est celle du suspect (note : c'est l'hypothèse de la poursuite). L'hypothèse 2 (H2) est que la voix sur l'échantillon n'est pas celle du suspect (note : c'est l'hypothèse de la défense).

La probabilité d'obtenir le résultat de l'analyse (E) avec l'hypothèse 1 est 100 fois plus élevée que la probabilité d'obtenir ce résultat avec l'hypothèse 2. Le rapport de vraisemblance est de 100/1. En d'autres mots, si on réutilise le tableau d'équivalence, on peut dire que nos résultats supportent l'hypothèse 1 de façon modérée à forte.

10.4.2.5 Questions sur les données d'apprentissage

- À partir de quelles données a-t-on entraîné le système pour évaluer l'hypothèse 1 et l'hypothèse 2?

10.4.2.6 Réponses

Nous avons entraîné le système à partir des caractéristiques connues du suspect, c'est-à-dire un homme dans la quarantaine et d'expression québécoise. Pour estimer P(E|H1), nous avons entraîné le système sur la voix sur l'échantillon 1. Pour estimer P(E|H2), nous avons entraîné le système sur une population de référence constituée de 30 voix qui sont fournies avec le système.

10.4.3 Fiabilité de l'explication

- Est-ce que le système produit une explication du résultat?
- D'autres extraits de l'appel au 9-1-1 ont-ils été comparés à l'enregistrement deux (pour confirmer la présence du suspect lors de l'appel au 9-1-1)?
- Le résultat a-t-il été confirmé par un outil distinct ou une autre technique (par exemple une analyse phonétique-acoustique)?

10.4.3.1 Réponses

Le système ne fournit pas d'explication. Nous n'avons pas fait d'analyses sur d'autres extraits de l'appel au 9-1-1 ni avec d'autres outils. Il n'y a pas eu d'autres types d'expertise.

10.4.4 Questions sur la performance

- Quelle est la performance générale du système

- Quelle est la performance pour les locuteurs masculins avec le même accent que le suspect?
- Le système est-il calibré pour favoriser les faux positifs ou faux négatifs?
- Quels sont facteurs qui peuvent avoir une incidence négative sur la performance du système?
- Lesquels de ces facteurs sont présents dans le cas d'espèce?
- Quelles sont les méthodes appliquées pour compenser ces facteurs?

10.4.4.1 Réponse

La performance du système est mesurée en termes de faux positifs et taux de faux négatifs. Nous calibrons le système pour que les deux taux soient identiques.

Dans des conditions optimales, le total des erreurs de type faux positifs et faux négatifs ne dépasse pas 2% et ce, pour l'ensemble des groupes démographiques qui sont supportés par le système, dont les locuteurs masculins à l'accent québécois.

On considère plusieurs facteurs de variabilité dans la voix. La variabilité interlocuteurs inclut les différences morphologiques, physiologiques et de prononciation entre les individus. C'est cette variabilité qui est exploitée par le système pour reconnaître le locuteur. La variabilité intralocuteur est dépendante de l'état physique et psychologique du locuteur et la variabilité intersession concerne les canaux d'enregistrement et de transmission. Ce sont les variabilités intralocuteurs et intersessions qui posent des difficultés dans la reconnaissance du locuteur. Dans le cas qui nous concerne, les facteurs suivants sont susceptibles d'augmenter le taux d'erreurs : la courte durée de l'enregistrement 1 et sa qualité moyenne en général, l'état psychologique probablement distinct du locuteur sur les deux enregistrements, les 5 ans qui se sont écoulés entre les deux enregistrements car la voix se modifie avec l'âge. Il se peut aussi que le suspect ait modifié volontairement sa voix lors de l'entrevue qui a donné lieu au deuxième enregistrement.

Le système normalise automatiquement les enregistrements pour atténuer la variabilité inter-sessions. De plus, pour atténuer les effets de la variabilité intra-locuteur et compenser le peu de phonèmes de l'enregistrement 1, le système a adapté le modèle statistique à un « modèle du monde » qui est un modèle standardisé représentant une grande variété de voix.

10.4.5 Questions sur les tests

- Quels tests ont été effectués sur des données qui présentent des caractéristiques similaires au cas d'espèce?
- Quelle est la taille des échantillons de test?
- Les tests ont-ils été validés et par qui?
- Les tests peuvent-ils être répétés (mêmes résultats avec d'autres données) et reproduits (mêmes résultats avec des données identiques)?

10.4.5.1 Réponses

Nous avons testé les performances du système en simulant les conditions techniques de prélèvements de l'enregistrement 1 ainsi que les bruits de fond. Nos résultats sont de 25% d'erreurs sur un échantillon de trente voix distinctes d'hommes dans la quarantaine avec un accent québécois. Ces trente voix sont différentes de celles utilisées pour l'apprentissage.

Les tests peuvent être reproduits facilement car nous conservons l'ensemble des cas de tests, les paramètres de système et les filtres sonores permettant de reproduire les conditions mentionnées. Les tests ont été validés par un collègue à l'interne.

10.4.6 Conclusions sur l'admissibilité de la preuve

Les données sources, soient les enregistrements, sont cohérentes avec l'objectif du système. Cependant, les données de l'échantillon 1 sont de qualité moyenne selon l'expert, entre autres à cause des bruits de fond et de la distance entre le micro et le

locuteur. De plus, le modèle produit avec l'échantillon 1, à cause de la courte durée de trois secondes[720], a dû être bonifié à l'aide d'un « modèle du monde » générique et donc non spécifique aux caractéristiques du locuteur. Il ne nous semble pas clair à quel point le modèle final est fidèle aux caractéristiques du locuteur. L'échantillon 1 doit être considéré comme incomplet. Qui plus est, il été a été pris il y 5 ans, durée au cours de laquelle la voix peut changer. Quant à l'enregistrement 2, étant donné que le locuteur connaissait l'objectif de l'entrevue, il a pu fausser sa voix et ajoute aux risques d'inexactitude de l'analyse. En somme, les données sources sont de piètre qualité.

La piètre qualité des enregistrements constitue la principale limite technique reconnue du système et pose un risque élevé d'engendrer des résultats inexacts. Le taux d'erreurs dans les conditions cumulées du cas d'espèce n'a pas été déterminé. On pourrait présumer que ce taux d'erreurs, en additionnant l'ensemble des facteurs de variabilités tels que l'âge et l'état psychologique du locuteur, soit supérieur à 25%, ce qui nous paraît élevé dans le contexte.

La population de référence sur lequel le système a été entraîné contient 30 voix d'hommes québécois dans la quarantaine. Ce nombre nous apparaît insuffisant pour refléter toute la variété et la richesse des voix d'une population de référence[721]. La

---

[720] Morrison, *supra* note 719 à la p 8. Trois secondes est la durée minimale acceptée par plusieurs systèmes.

[721] Cette étude démontre l'impact sur la performance de 30 cas versus 100 cas au niveau de la population de référence : David van der Vloed, « Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01) » (2016) 85 Speech Communication; *R v Slade & Others*, *supra* note 718 au para 178. Cette décision au Royaume-Uni abonde dans le même sens au sujet de la diversité de la population de référence.

courte durée de l'échantillon 1 n'a pas permis un entraînement à partir de l'échantillon lui-même. Pour ces raisons, les données d'entraînement nous paraissent inadéquates.

Sur le plan des tests, un volume de 30 cas de tests nous paraît insuffisant pour refléter l'ensemble des variétés morphologiques des locuteurs[722]. De plus les tests ne rendent pas compte des conditions réelles du cas d'espèce en termes de l'état psychologique, des cris sur l'enregistrement 1 et du passage du temps sur la voix.

Les régionalismes peuvent influencer l'accent. Or le système a été entraîné et testé sur des accents « québécois », ce qui ne rend pas compte des spécificités régionales. Il aurait été approprié d'utiliser la région précise du suspect dans l'analyse.

Le système commercial pour réaliser l'analyse est très répandu parmi les forces de l'ordre à l'international. Cependant, la validité scientifique de la reconnaissance vocale est contestée par la communauté scientifique et fait même l'objet d'un moratoire sur les expertises juridiques en France. Le fait qu'aucune certification et aucun standard n'encadre le développement de ces outils n'ajoute aucune garantie de fiabilité.

En somme, la fiabilité du résultat produit par le système d'apprentissage automatique et soumis en preuve n'a pas été démontrée dans le cas d'espèce. La valeur probante de la preuve, soit sa capacité à établir la vérité par sa force de conviction, son objectivité, sa qualité et son absence d'erreurs[723], est *faible*. En contrepartie, le rapport de vraisemblance produit par le système et supportant la thèse que le suspect ait

---

[722] Morrison, *supra* note 719 aux pp 30-31. L'auteur indique que le manque de représentativité des données dans l'affaire *R v Slade*, incluant les 27 cas de tests, ont contribué au rejet de la preuve.

[723] André Émond, *supra* note 63 à la p 82.

prononcé la phrase incriminante et qui incite dès lors à conclure à la présence du suspect sur les lieux du crime, risque d'orienter faussement le tribunal et ainsi détruire la présomption d'innocence du suspect. Les risques d'effets néfastes de la preuve sont *élevés* en regard de la valeur probante.

La Cour rejette donc la preuve de reconnaissance du locuteur à l'aide de l'outil d'apprentissage automatique.

CHAPITRE XI

RÉSUMÉ DES CONDITIONS FAVORABLES À LA FIABILITÉ

Les conditions favorables à la fiabilité de la preuve produite par un outil d'apprentissage automatique sont les suivantes :

- La population de référence à laquelle les statistiques s'appliquent est pertinente;
- Le modèle est individualisé selon les caractéristiques de la personne visée par le résultat;
- L'objectif du système est pertinent;
- La vision encodée du phénomène impliqué est généralement acceptée;
- Les données de tests et les données d'apprentissage sont complètes et représentatives;
- Les données impliquées dans l'obtention du résultat sont exactes et pertinentes;
- Les corrélations entre les données impliquées dans le résultat et le résultat sont cohérents;
- L'explication est fidèle au modèle et permet l'évaluation de divers scénarios;
- Les intervenants humains sont adéquatement formés sur les limites et contraintes du système;
- Les tests sont reproductibles et répétables;
- Le système est sécuritaire et robuste;
- Les systèmes autonomes font l'objet de procédures de tests adéquates;
- La performance du système est adéquate, tant globalement que pour les sous-groupes d'intérêt;
- Le code et la documentation du système, celle-ci reflétant l'ensemble des décisions prises lors de la conception, ont été révisés par les pairs;
- Le système est conforme aux normes de qualité généralement acceptées.

CONCLUSION

Un résultat d'analyse produite par une nouvelle technique doit être *fiable* pour être admissible en preuve au tribunal. En droit canadien, la fiabilité implique que le principe fondamental sur lequel repose la technique soit fiable et que les méthodes utilisées pour parvenir au résultat répondent à la méthode scientifique. L'explication du résultat doit être satisfaisante. La fiabilité peut être évaluée en fonction des critères énumérés dans *J.-L.J.*[724] soit, elle doit avoir fait l'objet de tests, d'une acceptation générale et d'une revue par les pairs et son taux d'erreur doit être explicite.

Cet ouvrage se veut un guide pour faciliter l'évaluation de la fiabilité d'un système d'apprentissage automatique selon la grille d'analyse ci-haut. Ainsi, tout en décrivant les composantes et le fonctionnement d'un tel système, nous avons identifié les défis techniques susceptibles d'impacter sa fiabilité. Ces défis engendrent des questions qui pourront être adressées au témoin expert lors de la présentation de la preuve. Les questions aideront à évaluer la fiabilité de la preuve et jauger sa force probante.

À la lecture de notre ouvrage, on pourra se questionner sur le rôle des tribunaux dans l'évaluation de la fiabilité d'un système d'apprentissage automatique. On peut identifier trois raisons qui incitent à questionner ce rôle : l'unicité de chaque système

---

[724] *J.-L.J.*, *supra* note 30.

dans un contexte précis, les connaissances requises et l'effort encouru pour évaluer la fiabilité des systèmes.

L'unicité des systèmes et du contexte. Notre ouvrage est limité aux aspects génériques de l'apprentissage automatique alors qu'une situation réelle impose l'évaluation de la fiabilité d'un système spécifique pour une affaire spécifique. En effet, chaque système d'apprentissage automatique, même s'il vise le même objectif, par exemple identifier un individu par la reconnaissance faciale est unique car il est le produit d'une myriade de décisions de la part des parties prenantes du système telles que les personnes qui conçoivent et utilisent le système. Chaque système doit donc être évalué en soi. La cour pourra se fier à la jurisprudence seulement si le système et sa version sont identiques au cas d'espèce. Concernant la version, le système aura pu évoluer avec l'ajout de nouvelles fonctionnalités ou des corrections d'erreurs. L'évaluation de sa fiabilité sera à refaire, au moins partiellement. Même si le système et sa version lors d'un cas de jurisprudence sont identiques au cas d'espèce, l'évaluation de la fiabilité devra être refaite en partie car certains aspects dépendent du contexte précis, par exemple les aspects d'individualisation, l'analyse de performance selon le groupe démographique et l'évaluation des échantillons de tests selon leur correspondance au cas d'espèce. Autrement dit, plus souvent qu'autrement, la cour devra évaluer ou réévaluer la fiabilité de chacune des preuves soumises en totalité ou en partie.

Les connaissances requises. Pour évaluer la fiabilité d'un système, l'ensemble des procédures, de la conception du système à la production du résultat final, doivent être évaluées. De plus, chaque domaine d'application, par exemple la reconnaissance du locuteur ou la reconnaissance faciale, comporte ses propres enjeux qui baliseront les limites et contraintes d'un système d'apprentissage automatique. En reconnaissance du locuteur par exemple, la variabilité de l'état psychologique du locuteur influencera la performance du système. En somme, l'évaluation de la fiabilité de la preuve en vue

d'établir son admissibilité en cour implique d'aborder le domaine d'application lui-même (par exemple la reconnaissance du locuteur) et l'ensemble des procédures concernant le système d'apprentissage automatique.

Dans une affaire en droit criminel, l'évaluation devra être des plus minutieuses étant donné l'impact sur les personnes concernées et ce, d'autant plus si la preuve rapproche le juge des faits de la question fondamentale de la culpabilité ou l'innocence de l'accusé[725].

Or, si l'on approuve la prémisse qu'un procès d'experts ne devrait pas se substituer à un procès devant juge et jury[726], le fait que le tribunal mène le témoignage d'expert sur la fiabilité d'un système d'apprentissage automatique comporte un risque de minimiser l'incertitude de la preuve. En effet, la propension naturelle de l'humain à faire confiance à la machine (le biais d'automation)[727], les biais cognitifs[728] et défis sémantiques[729] liés à l'interprétation des statistiques, et la somme des connaissances requises pour faire ressortir du témoignage expert les incertitudes du système et de l'explication algorithmique, posent un risque de minimiser l'incertitude de la preuve et donc surestimer sa valeur probante. Or, l'incertitude de la preuve contribue au doute

---

[725] *Ibid* au para 37.

[726] *White Burgess*, *supra* note 45 au para 18.

[727] Deeks, *supra* note 286.

[728] Kahneman et Tversky, *supra* note 225.

[729] Crispino, Muehlethaler et Cadola, *supra* note 233.

raisonnable et doit favoriser l'accusé[730]. Sachant que les forces de l'ordre ont plutôt tendance à se servir d'un système d'apprentissage automatique pour incriminer un suspect[731], minimiser l'incertitude du système défavorise le suspect.

Effort encouru. Faisant fi de la prémisse susmentionnée selon laquelle un procès d'experts ne devrait pas se substituer à un procès devant juge et jury, le juge des faits pourra exiger auprès d'un tiers indépendant des expertises supplémentaires, au-delà du témoignage expert, telles qu'une revue de code, une analyse de la documentation technique, une évaluation des jeux de tests, etc., et ce afin de mieux jauger la fiabilité de la technique. Or ces analyses, indispensables à l'évaluation minutieuse de la fiabilité[732], ont un coût et entraînent des délais. On se rappellera que l'admissibilité d'une preuve doit faire l'objet d'une analyse de sa valeur probante en regard de ses effets préjudiciables, lesquels incluent les coûts et délais[733]. Considérant l'ampleur de la tâche, ces effets préjudiciables pénaliseront le suspect, sachant que les forces de l'ordre ont plutôt tendance à l'incriminer et que ce dernier devra se défendre contre le résultat défavorable du système d'apprentissage automatique[734].

---

[730] *R c Oakes*, [1986] 1 RCS 103, 1986 CanLII 46 . La présomption d'innocence, au coeur du système criminel, implique que « la culpabilité soit établie hors de tout doute raisonnable [et] que ce soit à l'état qu'incombe la charge de la preuve ».

[731] Roth, *supra* note 421 à la p 3. Les erreurs de type faux positifs sont favorisées.

[732] CE, Commission, *supra* note 549. La proposition du Parlement européen exige ces analyses.

[733] *Mohan*, *supra* note 26.

[734] Roth, *supra* note 421 à la p 3.

On peut toutefois esquisser une piste de solutions à ces difficultés. Le rôle d'évaluation de la fiabilité générale d'un système d'apprentissage automatique plutôt que d'incomber à la cour, pourrait être confié par le législateur à un organe de certification neutre, impartial et possédant toute l'expertise nécessaire. Le Canada, se faisant, pourrait prendre exemple sur la proposition de la Commission européenne qui inclut une telle certification[735]. Le Canada pourrait ainsi définir des standards pour l'usage de preuves découlant d'analyses produites par des systèmes d'apprentissage automatique dans les affaires pénales et se doter d'un organisme indépendant de certification. Cette certification de systèmes d'apprentissage automatique à des fins de preuve, en amont de la demande d'admission en preuve, pourrait alléger le processus judiciaire tout en assurant que la fiabilité générale du système serait analysée avec toute l'expertise nécessaire. Ainsi, le législateur pourrait faire naître de la certification de conformité d'un système d'apprentissage automatique une présomption selon laquelle le système remplit les conditions générales de fiabilité. Une présomption semblable existe à l'égard des éthylomètres approuvés par le procureur général du Canada en vertu des articles 320.31(1) et 320.39 du *Code criminel*[736]. Seuls les aspects contextuels plus spécifiques nécessiteraient alors d'être analysés en termes de fiabilité par les tribunaux. Cela faciliterait le travail de la cour sans usurper son rôle du juge des faits. Aussi, cela réduirait le poids que fait peser sur les épaules de la défense le besoin de procéder à toute l'analyse complexe à laquelle nous avons fait référence dans le cadre

---

[735] CE, Commission, *supra* note 549 art 15 et 19.

[736] *Code criminel*, LRC 1985, c C-46. Concernant les conditions de validité d'une telle présomption et des moyens de la renverser dans une perspective de protection de la présomption d'innocence, voir *R c St-Onge Lamoureux*, 2012 CSC 57, [2012] 3 RCS 187; et *R c Cyr-Langlois*, 2018 CSC 54, [2018] 3 RCS 456 qui analysent l'article qu'est venu remplacer le récent art. 320.31(1) du C. cr.

de cet ouvrage pour jauger la force d'une preuve algorithmique; celles issues d'un système ne passant pas la barre d'un certain niveau de fiabilité seraient exclues d'office.

Cette piste de solution n'est sans doute pas parfaite et nécessite une réflexion approfondie. Il va sans dire, par exemple, que la distinction entre la fiabilité générale et celle découlant des éléments contextuels devrait faire l'objet de précisions. Toutefois, que nous options pour la création ou non d'un modèle de certification ayant pour effet de créer une présomption de fiabilité, l'apprentissage automatique est appelé à poser des défis grandissants aux juristes qui ne sont pas férus de technologies. L'acquisition d'une certaine littératie technologique, déjà reconnue comme obligation déontologique[737], se fera des plus pressantes pour les criminalistes.

---

[737] *Code de déontologie des avocats*, RLRQ c B-1, r. 3.1, art 21.:

> L'avocat exerce avec compétence ses activités professionnelles. À cette fin, il développe et tient à jour ses connaissances et ses habiletés.
>
> Pour l'application du premier alinéa, font partie des connaissances et des habiletés que l'avocat développe et tient à jour celles relatives aux technologies de l'information qu'il utilise dans le cadre de ses activités professionnelles.

BIBLIOGRAPHIE

INSTRUMENTS INTERNATIONAUX

OCDE, Instruments juridiques, *Recommandation du Conseil sur l'intelligence artificielle,* no de doc OECD/LEGAL/0449, 2019.

INSTRUMENTS RÉGIONAUX

LÉGISLATION

Europe

CE, *Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil*, [2016], JO, L119/89.

CE, *Résolution du Parlement européen du 6 octobre 2021 sur l'intelligence artificielle en droit pénal et son utilisation par les autorités policières et judiciaires dans les affaires pénales*, 2021.

CE, *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données),* [2016], JO, L119/1 [RGPD].

CE, Commission, *Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen, et au Comité des régions: Façonner l'avenir numérique de l'Europe,* Bruxelles, 2020.

Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union 2021/0106 (COD), 2021.

Code des relations entre le public et l'administration.

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (1), JO, 8 octobre 2016.

INSTRUMENTS NATIONAUX

LÉGISLATION

Canada

Canada, Conseil du Trésor, *Directive sur la prise de décision automatisée*, 2019.
Charte canadienne des droits et libertés, partie I de la *Loi constitutionnelle de 1982*, constituant l'annexe B de la *Loi de 1982 sur le Canada* (R-U), 1982, c 11.

Code criminel, LRC 1985, c C-46.

Code de déontologie des avocats, RLRQ c B-1, r. 3.1.

Code civil du Québec.

Loi concernant le cadre juridique des technologies de l'information, RLRQ, c C-1.1.

Loi modernisant des dispositions législatives en matière de protection des renseignements personnels, LQ 2021, c 25, 2021.

La Loi sur la protection des renseignements personnels et les documents électroniques, LC 2000, c 5.

Loi sur la preuve au Canada, LRC 1985, c C-5.

Loi sur la protection des renseignements personnels, LRC 1985, c P-21.

États-Unis

E-Government Act of 2002, Pub L No 107-347, 116 Stat 2899.

Equal Credit Opportunity Act, 15 USC 1691 (2011).

Federal Rules of Evidence, 28 USC (2022).

Freedom of Information Act of 1966, Pub L No 89-487, 80 Stat 250.

Privacy Act, 5 USC § 552a (1974).

É-U, Bill HR 2231, *Algorithmic Accountability Act*, 116e Cong, 2019.

É-U, Bill S 3284, *Ethical Use of Facial Recognition Act,* 116e Cong, 2020.

JURISPRUDENCE

Canada

Barendregt c Grebliunas, [2022] CSC 22.

Commission scolaire de Victoriaville c La Reine, [2002] CanLII 61082.

Ewert c Canada, [2018] 2018 CSC 30, [2018] 2 RCS 165.

Graat c La Reine, [1982] 2 RCS 81.

Hotel central (Victoriaville) inc Ltée c Compagnie d'assurance reliance Ltée, [1998] 1998 CanLII 12934 (QC CA).

Imperial Tobacco Canada ltée c Conseil québécois sur le tabac et la santé, [2019] QCCA 358.

R c Abbey, [1982] 2 RCS 24, 1982 CanLII 25.

R c Bingley, 2017 CSC 12, [2017] 1 RCS 170.

R c Cyr-Langlois, 2018 CSC 54, [2018] 3 RCS 456.

R c J-LJ, [2000] 2000 CSC 51, [2000] 2 RCS 600 .

R c Mohan, [1994] 2 RCS 9, 1994 CanLII 80.

R c Morin, [1988] 2 RCS 345, 1988 CanLII 8.

R c Oakes, [1986] 1 RCS 103, 1986 CanLII 46.

R c St-Onge Lamoureux, 2012 CSC 57, [2012] 3 RCS 187.

R c Trochym, 2007 CSC 6, [2007] 1 RCS 239.

R v Abbey, [2009] 2009 ONCA 624, 97 OR (3e) 330.

R v Béland, [1987] [1987] 2 SCR 398, 1987 CanLII 27.

R v Dosanjh, [2019] 2019 ONSC 1320.

R v Johnston, [1992] 1992 CanLII 12790 (ON SC).

R v Medvedew, 91 DLR (3d) 21, 6 CR (3d) 185, 1978 CanLII 2101.

R v Melaragni, 1992 CanLII 12764 (ON SC).

R v Pinch, [2011] 2011 ONSC 5484.

Withler c Canada (Procureur général), [2011] CSC 12.

White Burgess Langille Inman Ltd c Abbott and Haliburton Co Ltd, 2015 CSC 23, [2015] 2 RCS 182.

États-Unis

Daubert v Merrell Dow Pharmaceuticals Inc, 509 US 579, 1993 US Lexis 4408.

Frye v United States, 293 F 1013, 1923 US App Lexis 1712.

R v Slade & Others, [2015] EWCA Crim 71.

State of New Jersey v Corey Pickett, [2021] 466 NJ Super 270, 2021 NJ Super Lexis 17.

State v Loomis, [2016] 2016 WI 68, 2016 Wisc Lexis 178.

United States v Baller, 519 F (2d) 463, 1975 US App Lexis 13781.

Autre

R (Bridges) v CCSWP and SSHD, [2019] EWHC 2341 (Admin) .

Cons const, 12 juin 2018, [2018], 2018-765 DC.

DOCUMENTS GOUVERNEMENTAUX

Canada

Canada, Commissariat à la protection de la vie privée du Canada, « Clearview AI cesse d'offrir sa technologie de reconnaissance faciale au Canada », (6 juillet 2020), en ligne: <https://www.priv.gc.ca/fr/nouvelles-du-commissariat/nouvelles-et-annonces/2020/nr-c_200706/>.

Canada, Commissariat à la protection de la vie privée, « Survol de la Loi sur la protection des renseignements personnels », (23 août 2019), en ligne: <https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/la-loi-sur-la-protection-des-renseignements-personnels/pa_brief/>.

Canada, Commissariat à la protection de la vie privée, « Survol de la LPRPDE », (1 mai 2019), en ligne: <https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/la-loi-sur-la-protection-des-renseignements-personnels-et-les-documents-electroniques-lprpde/lprpde_survol/>.

Canada, Commissariat à la protection de la vie privée, « Un cadre réglementaire pour l'IA : recommandations pour la réforme de la LPRPDE », (12 novembre 2020), en ligne: <https://www.priv.gc.ca/fr/a-propos-du-commissariat/ce-que-nous-faisons/consultations/consultations-terminees/consultation-ai/reg-fw_202011/#fn10>.

Canada, Commission de réforme du droit, *Les méthodes d'investigation scientifique*, document de travail 34, 1984.

Canada, Institut national de la magistrature*, Manuel scientifique à l'intention des juges canadiens,* 2018 [Manuel scientifique].

Canada, Ministère de la Justice, « Automatisation de la justice - Tendances en matière de justice 2 : Automatisation de la justice. Un aperçu de l'avenir des technologies dans

le système judiciaire », (2 avril 2008), en ligne: *Gouvernement du Canada* <https://www.justice.gc.ca/fra/pr-rp/jr/tmj2-jt2/p3.html>.

Canada, Secrétariat du Conseil du Trésor du, « Outil d'évaluation de l'incidence algorithmique », (22 mars 2021), en ligne: <https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html>.

Québec, Commission des droits de la personne et des droits de la jeunesse*, Mémoire à la Commission d'accès à l'information sur le document de consultation «Intelligence artificielle»*, 2020.

Québec, Commission d'accès à l'information du Québec, « Biométrie », (mars 2021), en ligne: <https://www.cai.gouv.qc.ca/biometrie/>.

États-Unis

États-Unis, President's Council of Advisors on Science and Technology*, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, 2016.

Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection Version 2.2*, 2017.

National Institute of Standards and Technology, *Face Recognition Vendor Test (FRVT) Part 2: Identification, NIST IR 8271 Draft Supplement*, 2021.

National Institute of Standards and Technology, *Face Recognition Vendor Test Part 3: Demographic Effects*, NIST IR 8280, 2019.

Europe

CE, *Lignes directrices en matière d'éthique pour une IA digne de confiance: Groupe d'experts de haut niveau sur l'intelligence artificielle*, 2019.

France, Groupe de travail «article 29» sur la protection des données, *Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679*, 2018.

OUVRAGES COLLECTIFS

Bellemare, Nicolas, « Chapitre III - Les procédures précédant le procès en matière criminelle » dans *Droit pénal: procédure et preuve* Collection de droit 2019-2020, Montréal, Québec, Éditions Yvon Blais, 2019, 41.

« Chapitre VII - La preuve pénale » dans *Droit pénal: procédure et preuve,* vol 12, Montréal, Éditions Yvon Blais, 2019.

Bonastre, Jean-François, « 1990-2020 : retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire » dans Gilles Adda, Maxime Amblard et Karën Fort, dir, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, Nancy, France, ATALA, 2020, 38.

« Forensic Speaker Recognition: Mirages and Reality » dans *Individual Differences in Speech Production and Perception*, Peter Lang International Academic Publishers, 2015, 255.

Encinas De Munagorri, Rafael, « Les problèmes de preuve posés par l'évolution des sciences et des technologies » dans *Applied Ethics at the Turn of Millenium*, 2001.

Gundersen, Odd Erik et Sigbjørn Kjensmo, « State of the Art: Reproducibility in Artificial Intelligence » dans *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kluttz, Daniel, Kohli, Nitin, et Mulligan, Deirdre K, « Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions » dans *After the Digital Tornado* Cambridge University Press, Kevin Werbach, 2020, 137.

Morrison, Geoffrey Stewart, Cuiling Zhang et Ewald Enzinger, « Forensic speech science » dans Ian Freckelton et Hugh Selby, dir, *Expert Evidence*, Sydney, Australia, 2019.

Patenaude, Pierre, « De Mohan à J.-L.J., de Daubert à Khumo: qu'en est-il de la preuve scientifique ou technique innovatrice? » dans *Développements récents en droit administratif et constitutionnel 2002*, Cowansville, Yvon Blais, 2002, 111.

Roth, Andrea, « The Use of Algorithms in Criminal Adjudication » dans *The Cambridge Handbook of the Law of Algorithms* Cambridge Law Handbooks, Cambridge University Press, 2020, 407.

MONOGRAPHIES

André Émond, *Introduction au droit canadien,* 2e éd, Montréal, Wilson & Lafleur, 2016.

Barocas, Solon, Moritz Hardt et Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019.

Bollen, Kenneth et al, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*, Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences, National Science Foundation, 2015.

Cofone, Ignacio, *Propositions stratégiques aux fins de la réforme de la LPRPDE élaborées en réponse au rapport sur l'intelligence artificielle*, Commissariat à la protection de la vie privée du Canada, 2020.

David Hume, *Enquiry on Human Understanding*, oxford university press éd, 2007.

Doyle, Ken, *Bayes Theorem. Can Statistics Help Guide a Verdict in the Courtroom?*, The ISHI Report, News from the World of DNA Forensics, International Symposium on Human Identification, 2021.

Ducharme, Léo, *Précis de la preuve*, 6e éd, Montréal, Wilson & Lafleur, 2005.

Eubanks, Virginia, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, USA, St. Martin's Press, Inc., 2018.

Fjeld, Jessica et al, *Principled Artificial Intelligence : Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, 2020.

Fuster, Dr Gloria Gonzalez, *Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights*, Département thématique des droits des citoyens et des affaires constitutionnelles du Parlement européen, 2020.

Karpf, Anne, *La voix. Un univers invisible*, Hors collection, Paris, Autrement, 2008.

Molnar, Petra et Lex Gill, *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System*, International Human Rights Program, Faculty of Law, University of Toronto, 2018.

Murphy, Kevin P, *Machine learning - A Probabilistic Perspective*, Adaptive computation and machine learning series, Cambridge, MA, MIT Press, 2012.

O'Neil, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New Tork, Crown Publishing Group, 2016.

Paciocco, David M, *The Law of Evidence*, 8e éd, Toronto, Ontario, Irwin Law, 2020.

Robertson, Kate, Cynthia Khoo et Yolanda Song, *To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada*, Citizen Lab, University of Toronto, 2020.

Russell, Stuart et Peter Norvig, *Intelligence artificielle: Avec plus de 500 exercices*, Pearson Education France, 2010.

ARTICLES DE REVUES SCIENTIFIQUES

Adadi, A et M Berrada, « Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) » (2018) 6 IEEE Access 52138.
Babic, Boris et al, « Beware explanations from AI in health care » (2021) 373:6552 Science 284.

Baker, Monya, « 1,500 Scientists Lift the Lid on Reproducibility » (2016) 533:7604 Nature 452.

Bansal, Gagan et al, « Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance » (2019) 7:1 Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 2.

Beaudouin, Valérie et al, « Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach » (2020) arXiv:200307703 [cs], en ligne: <http://arxiv.org/abs/2003.07703>.

Bellemare, Daniel A, « L'identification d'un accusé par ses vocogrammes » (1978) 56:3 R du B can 440.

Binnie, Ian, « Science in the Courtroom: The Mouse That Roared » (2007) 56 UNBLJ 307.

Boë, Louis-Jean et Jean-François Bonastre, « L'identification du locuteur : 20 ans de témoignage dans les cours de Justice. Le cas du Lipsadon » (2012) Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP 417-424.

Bourque, Jimmy, Jean-Guy Blais et François Larose, « L'interprétation des tests d'hypothèses : p, la taille de l'effet et la puissance » (2009) 35:1 Revue des sciences de l'éducation 211.

Brown, Russell, « The Possibility of "Inference Causation": Inferring Cause-in-Fact and the Nature of Legal Fact-Finding » (2010) 55:1 RD McGill 1.

Buolamwini, Joy et Timnit Gebru, « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification » (2018) 81 Proceedings of Machine Learning Research 1-15.

Burrell, Jenna, « How the Machine "Thinks" : Understanding Opacity in Machine Learning Algorithms » (2016) 3:1 Big Data & Society 2053951715622512.

Busino, Giovanni, « La preuve dans les sciences sociales » (2003) 41:128 Revue européenne des sciences sociales 11.

Bzdok, Danilo, « Classical Statistics and Statistical Learning in Imaging Neuroscience » (2017) 11 Frontiers in Neuroscience 1.

Bzdok, Danilo, Naomi Altman et Martin Krzywinski, « Statistics versus machine learning » (2018) 15:4 Nature Methods 233.

Carriquiry, Alicia et al, « Machine learning in forensic applications » (2019) 16 Significance (Royal Statistical Society) 29-35.

Crispino, Frank, Cyril Muehlethaler et Liv Cadola, « Vraisemblable n'est pas probable. De la nécessité d'une sémantique rigoureuse entre scientifiques et juristes » (2020) 2:1 Revue canadienne de justice et droit 95.

Deeks, Ashley, « The Judicial Demand for Explainable Artificial Intelligence, Virginia Public Law and Legal Theory Research Paper 2019-51 » (2019) SSRN, en ligne: <https://papers.ssrn.com/abstract=3440723>.

Doran, Derek, Sarah Schulz et Tarek R Besold, « What Does Explainable AI Really Mean? A New Conceptualization of Perspectives » (2017) arXiv:171000794 [cs], en ligne: <http://arxiv.org/abs/1710.00794>.

Doshi-Velez, Finale et al, « Accountability of AI Under the Law: The Role of Explanation » (2019) arXiv:171101134 [cs, stat], en ligne: <http://arxiv.org/abs/1711.01134>.

Doshi-Velez, Finale et Been Kim, « Towards A Rigorous Science of Interpretable Machine Learning » (2017) arXiv:170208608 [cs, stat], en ligne: <http://arxiv.org/abs/1702.08608>.

Dror, Itiel, « The Ambition to be Scientific : Human Expert Performance and Objectivity » (2013) 53:2 Science & Justice: Journal of the Forensic Science Society 81-82.

Duval Hesler, Nicole, « L'admissibilité des nouvelles théories scientifiques » (2002) 62 R du B 359.

Eaglin, Jessica, « Constructing Recidivism Risk » (2017) 67:59 Emory LJ 60.

Encinas de Munagorri, Rafael, « La recevabilité d'une expertise scientifique aux États-Unis » (1999) 51:3 RIDC 621.

Ezer, Neta et al, « Trust Engineering for Human-AI Teams » (2019) 63 Proceedings of the Human Factors and Ergonomics Society Annual Meeting 322.

Felderer, Michael et Rudolf Ramler, « Quality Assurance for AI-based Systems: Overview and Challenges » (2021) arXiv:210205351 [cs], en ligne: <http://arxiv.org/abs/2102.05351>.

Feldman, Michael et al, « Certifying and removing disparate impact » (2015) arXiv:14123756 [cs, stat], en ligne: <http://arxiv.org/abs/1412.3756>.

Fenton, Norman, « Improve statistics in court » (2011) 479:7371 Nature 36.

Fenton, Norman et Martin Neil, « Avoiding Probabilistic Reasoning Fallacies in Legal Practice using Bayesian Networks » (2011) 36 Australian Journal of Legal Philosophy.

Ghosh, Pallab, « AAAS: Machine learning “causing science crisis” », *BBC News* (16 février 2019), en ligne: <https://www.bbc.com/news/science-environment-47267081>.

Gibney, Elizabeth, « This AI researcher is trying to ward off a reproducibility crisis » (2019) 577:7788 Nature 14.

Gilpin, Leilani H et al, « Explaining Explanations: An Overview of Interpretability of Machine Learning » (2019) arXiv:180600069 [cs, stat], en ligne: <http://arxiv.org/abs/1806.00069>.

Gold, Erica et Peter French, « International Practices in Forensic Speaker Comparisons: Second Survey » (2019) 26:1 International Journal of Speech, Language and the Law 1.

Goldberg, Richard, « Epidemiological Uncertainty, Causation, and Drug Product Liability » (2014) 59:4 RD McGill 777.

Greenberg, Craig S et al, « Two Decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology » (2020) Computer Speech and Language, en ligne: <https://www.nist.gov/publications/two-decades-speaker-recognition-evaluation-national-institute-standards-and-technology>.

Grimaud, Marie Angèle, « Les enjeux de la recevabilité de la preuve d’identification par ADN dans le système pénal canadien » (1994) 24:2 RDUS 293.

Guidotti, Riccardo et al, « A Survey of Methods for Explaining Black Box Models » (2018) arXiv:180201933 [cs], en ligne: <http://arxiv.org/abs/1802.01933>.

Gundersen, Odd Erik, Yolanda Gil et David Aha, « On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications » (2018) 39 AI Magazine 56.

Gunning, David et David Aha, « DARPA’s Explainable Artificial Intelligence (XAI) Program » (2019) 40:2 AI Magazine 44.

Hoffman, Robert R et al, « Metrics for Explainable AI: Challenges and Prospects » (2019) arXiv:181204608 [cs], en ligne: <http://arxiv.org/abs/1812.04608>.

Hong, Nicole, « Court to Rule on Voice Analysis in Terrorism Trial », *Wall Street Journal* (11 mai 2015), en ligne: <https://www.wsj.com/articles/court-to-rule-on-voice-analysis-in-terrorism-trial-1431361065>.

Jordan, M I et T M Mitchell, « Machine learning: Trends, perspectives, and prospects » (2015) 349:6245 Science 255.

Kahneman, Daniel et Amos Tversky, « On the psychology of prediction » (1973) 80:4 Psychological Review 237.

Katherine Kwong, « The Algorithm says you did it: the use of black box algorithms to analyze complex DNA evidence. » (2017) 31:1 Harvard Journal of Law & Technology 275.

Kleinberg, Jon et al, « Discrimination in the Age of Algorithms » (2018) 10 Journal of Legal Analysis 113.

Knight, Will, « If AI's So Smart, Why Can't It Grasp Cause and Effect? » *Wired* (3 septembre 2020), en ligne: <https://www.wired.com/story/ai-smart-cant-grasp-cause-effect/>.

Koenecke, Allison et al, « Racial disparities in automated speech recognition » (2020) 117:14 Proceedings of the National Academy of Sciences 7684-7689.

———, « Racial disparities in automated speech recognition » (2020) 117:14 Proceedings of the National Academy of Sciences 7684.

Leese, Matthias, Mareile Kaufmann et Simon Egbert, « Predictive Policing and the Politics of Patterns » (2019) 59 British Journal of Criminology 674.

Lehr, David et Paul Ohm, « Playing with the Data: What Legal Scholars Should Learn about Machine Learning » (2017) 51:2 UCD L Rev 653.

Limpert, P Brad, « Beyond the Rule in Mohan: A New Model for Assessing the Reliability of Scientific Evidence » (1996) 54:1 UT Fac L Rev 65.

Lindgren, Simon, « Hacking Social Science for the Age of Datafication » (2019) 1:1 Journal of Digital Social Research 1.

Lipton, Zachary C, « The Mythos of Model Interpretability » (2017) arXiv:160603490 [cs, stat], en ligne: <http://arxiv.org/abs/1606.03490>.

Lu, Yang, « Artificial Intelligence: A Survey on Evolution, Models, Applications and Future Trends » (2019) 6:1 Journal of Management Analytics 1.

Mann, Monique et Tobias Matzner, « Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination » (2019) 6:2 Big Data & Society 1.

Martinez, Frédéric, « L'individu face au risque : l'apport de Kahneman et Tversky » (2010) N° 161:3 Idees economiques et sociales 15.

Milan, Kishan, « Quality Assurance Factors: Key Aspects for Software Quality Control and Testing » (2021) 69 International Journal of Computer Trends and Technology.

Miller, George A, « The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information » (1956) 63:2 Psychological Review 81.

Miller, Tim, « Explanation in Artificial Intelligence: Insights from the Social Sciences » (2018) arXiv:170607269 [cs], en ligne: <http://arxiv.org/abs/1706.07269>.

Mitchell, Tom M, « Does Machine Learning Really Work? » (1997) 18:3 AI Magazine 11.

Morrison, Geoffrey Stewart, « Admissibility of forensic voice comparison testimony in England and Wales » (2018) 2018:1 Crim L Rev 20.

Nagarajan, Prabhat, Garrett Warnell et Peter Stone, « The Impact of Nondeterminism on Reproducibility in Deep Reinforcement Learning » (2018) International Conference on Machine Learning, en ligne: <https://openreview.net/forum?id=S1e-OsZ4e7>.

Nawaz, Nishad, « How Far Have We Come With The Study Of Artificial Intelligence For Recruitment Process » (2019) 8:7 International Journal of Scientific & Technology Research 488.

Nguyen, Anh, Jason Yosinski et Jeff Clune, « Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images » (2015) arXiv:14121897 [cs], en ligne: <http://arxiv.org/abs/1412.1897>.

Nutter, Patrick, « Machine Learning Evidence: Admissibility and Weight » (2019) 21:3 U Pa J Const L 919.

Paciocco, David, « L'évaluation du témoignage d'opinion pour en établir l'admissibilité : les leçons récentes du droit de la preuve » (1995) 26:3 RGD 425.

Patenaude, Pierre, « De l'expertise judiciaire dans le cadre du procès criminel et de la recherche de la vérité : quelques réflexions » (1997) 27:1 RDUS 1-47.

Patenaude, Pierre, « Modern Scientific Evidence » (2000) 30:2 RDUS 407-417.

Pégny, Maël et Issam Ibnouhsein, « Quelle transparence pour les algorithmes d'apprentissage machine ? » (2018) 32 Revue d'intelligence artificielle 447.

Polyzotis, Neoklis et al, « Data Management Challenges in Production Machine Learning » (2017) Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17) 1723-1726.

Ranschaert, Erik, « Artificial Intelligence in Radiology: Hype or Hope? » (2018) 102:S1 Journal of the Belgian Society of Radiology 20.

Rashidi, Hooman et al, « Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods » (2019) 6 Academic Pathology.

Remmers, Julius, « Legal Service by Automated Legal Software and Its Legal Impacts, in Particular under the German Act On Out-of-Court Legal Services » (2018) SSRN, en ligne: <https://papers.ssrn.com/abstract=3491347>.

Ribeiro, Marco Tulio, Sameer Singh et Carlos Guestrin, « "Why Should I Trust You?": Explaining the Predictions of Any Classifier » (2016) KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135-1144.

Richmond, Karen McGregor, « AI, Machine Learning, and International Criminal Investigations: The Lessons From Forensic Science » (2020) iCourts Working Paper Series, No 22, en ligne: <https://papers.ssrn.com/abstract=3727899>.

Roth, Andrea, « Machine Testimony » (2017) 126:7 Yale LJ, en ligne: <https://digitalcommons.law.yale.edu/ylj/vol126/iss7/1>.

Roth, Andrea L, « Trial by Machine » (2016) 104:5 Geo LJ 48.

Schmidt, Jonathan et al, « Recent advances and applications of machine learning in solid-state materials science » (2019) 5:1 npj Computational Materials 1.

Selbst, Andrew D et al, « Fairness and Abstraction in Sociotechnical Systems » (2018) 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*) 59.

Selbst, Andrew D et Solon Barocas, « The Intuitive Appeal of Explainable Machines » (2018) 87 Fordham L Rev 1085.

Sherrin, Christopher, « Earwitness Evidence: The Reliability of Voice Identifications » (2016) 52:3 Osgoode Hall LJ 819.

Solan, Lawrence et PM Tiersma, « Hearing voices: Speaker identification in court » (2003) 54:2 Hastings LJ 373.

Stettler, Gabriel, « L'administration de la preuve scientifique en droit Nord-Américain. » (2019) 97:1 Can Bar Rev 177.

Sugimura, Peter et Florian Hartl, « Building a Reproducible Machine Learning Pipeline » (2018) arXiv:181004570 [cs, stat], en ligne: <http://arxiv.org/abs/1810.04570>.

Swedloff, Rick, « The New Regulatory Imperative for Insurance » (2019) 61:6 Boston College L Rev 2033.

Varnava, Christiana, « Technology Takes the Wheel » (2019) 2:8 Nature Electronics (Nature Publishing Group) 319.

Vloed, David van der, « Evaluation of Batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01) » (2016) 85 Speech Communication.

Wachter, S, B Mittelstadt et L Floridi, « Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation » (2017) 7:2 International Data Privacy Law 76.

Wachter, Sandra, Brent Mittelstadt et Chris Russell, « Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR » (2018) 31:2 Harvard JL & Tech 841.

———, « Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI » (2021) 41 Computer Law & Security Review, en ligne: <https://www.sciencedirect.com/science/article/pii/S0267364921000406>.

Woods, Walt, Jack Chen et Christof Teuscher, « Adversarial Explanations for Understanding Image Classification Decisions and Improved Neural Network Robustness » (2019) 1:11 Nat Mach Intell 508.

Zheng, Linlin et al, « When Automatic Voice Disguise Meets Automatic Speaker Verification » (2021) 16 IEEE Transactions on Information Forensics and Security 824.

Khelif, Khaled et al, *SIIP: An Innovative Speaker Identification Approach for Law Enforcement Agencies*, présenté à Big Data and Artificial Intelligence for Military Decision Making, 2018.

MÉMOIRES ET THÈSES

Ajili, Moez, *Reliability of voice comparison for forensic applications*, thèse de doctorat en informatique, Université d'Avignon et des Pays de Vaucluse, 2017 [non publiée].

Collard, Bastien, *L'impact de l'intelligence artificielle dans la gestion de portefeuilles* (mémoire de M. Sc, Université de Louvain, 2019) [non publié].

Jourani, Reda, *Reconnaissance automatique du locuteur par des GMM à grande marge*, thèse de doctorat en informatique, Université Paul Sabatier - Toulouse III, 2012 [non publiée].

Mehra, Sidharth, *Detection of Offensive Language in Social Media Posts*, mémoire de maîtrise en intelligence artificielle au département d'informatique, Cork Institute of Technology, 2020 [non publiée].

RAPPORTS DE RECHERCHE

Céline Castets-Renard, Émilie Guiraud, et Jacinthe Avril-Gagnon, « Cadre juridique applicable à l'utilisation de la reconnaissance faciale par les forces de police dans

l'espace public au Québec et au Canada. Éléments de comparaison avec les États-Unis et l'Europe » (2020) Observatoire international sur les impacts sociétaux de l'IA et du numérique, Chaire de recherche IA responsable à l'échelle mondiale, en ligne: <https://www.docdroid.com/YIDTjrr/cadre-juridique-applicable-a-lutilisation-de-la-reconnaissance-faciale-par-les-forces-de-police-dans-lespace-public-au-quebec-et-au-canada-pdf>.

Partnership on AI, *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*, 2019.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, 2019.

Presser, Jill R. et Robertson, Kate, *AI Case Study: Probabilistic Genotyping DNA Tools in Canadian Criminal Courts*, Commission du droit de l'Ontario, 2021.

Université de Toronto, *Canada's AI Ecosystem: Government Investment Propels Private Sector Growth*, 2020.

Villani, Cédric et al, *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*, 2018.

Xianhong Hu et al, *Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective*, UNESCO 62530, 2019.

ENCYCLOPÉDIES

Altman, Andrew, « Discrimination » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

Angius, Nicola, Giuseppe Primiero et Raymond Turner, « The Philosophy of Computer Science » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2021 éd, Metaphysics Research Lab, Stanford University, 2021.

Bringsjord, Selmer et Naveen Sundar Govindarajulu, « Artificial Intelligence » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, summer 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

Copeland, BJ, « Artificial Intelligence » dans *Encyclopedia Britannica*, 2021.

Fidler, Fiona et John Wilcox, « Reproducibility of Scientific Results » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, summer 2021 éd, Metaphysics Research Lab, Stanford University, 2021.

Henderson, Leah, « The Problem of Induction » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

Hodges, Andrew, « Alan Turing » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2019 éd, Metaphysics Research Lab, Stanford University, 2019.

Romeijn, Jan-Willem, « Philosophy of Statistics » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2017 éd, Metaphysics Research Lab, Stanford University, 2017.

Schulte, Oliver, « Formal Learning Theory » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, spring 2018 éd, Metaphysics Research Lab, Stanford University, 2018.

The Editors of Encyclopaedia Britannica, « Causation » dans *Encyclopedia Britannica*, 2009.

Weirich, Paul, « Causal Decision Theory » dans Edward N Zalta, dir, *The Stanford Encyclopedia of Philosophy*, winter 2020 éd, Metaphysics Research Lab, Stanford University, 2020.

AUTRES

Aley-Raz, Almog, *«Voice Biometrics, The Silent Revolution, Commercial and Government success stories»,* Israel's Biometrics and Strong Authentification Conference, 2016.

Allen, Robin et Dee Masters, « French Parcoursup Decision », (16 avril 2020), en ligne: *AI Lawhub* <https://ai-lawhub.com/2020/04/16/french-parcoursup-decision/>.

Amazon, « Amazon.ca Aide: FAQ sur Alexa et les appareils Alexa », (novembre 2021), en ligne: *Amazon* <https://www.amazon.ca/-/fr/gp/help/customer/display.html?nodeId=201602230>.

Amini, Alexander, « Introduction to Deep Learning (6.S191) », (27 janvier 2020), en ligne: *MIT* <https://www.youtube.com/watch?v=njKP3FqW3Sk>.

Angwin, Julia et al, « Machine Bias », (23 mai 2016), en ligne: *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Apple, « Siri », (mai 2021), en ligne: *Apple (CA)* <https://www.apple.com/ca/fr/siri/>.

Athalye, Anish et al, « Fooling Neural Networks in the Physical World », (31 octobre 2017), en ligne: *LabSix* <https://www.labsix.org/physical-objects-that-fool-neural-nets/>.

Bajada, Josef, « Symbolic vs Connectionist A.I. », (9 avril 2019), en ligne: *Towards Data Science* <https://towardsdatascience.com/symbolic-vs-connectionist-a-i-8cf6b656927>.

Berkeley, « Artificial Intelligence: A Modern Approach, 4th US ed. », en ligne: *Berkeley* <http://aima.cs.berkeley.edu/>.

Blue J, « Predictive Tax and Employment Law Software », (2021), en ligne: <https://www.bluej.com/ca>.

Brownlee, Jason, « 14 Different Types of Learning in Machine Learning », (10 novembre 2019), en ligne: *Machine Learning Mastery* <https://machinelearningmastery.com/types-of-learning-in-machine-learning/>.

Brownlee, Jason, « Difference Between Algorithm and Model in Machine Learning », (28 avril 2020), en ligne: *Machine Learning Mastery* <https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/>.

Buolamwini, Joy, « Gender Shades », (2021), en ligne: *MIT Media Lab* <https://www.media.mit.edu/projects/gender-shades/overview/>.

Casetext, « Casetext Research: Best Legal Research Software », (2021), en ligne: *Casetext* <https://casetext.com/research/>.

Connely, Thomas, « Cambridge University law students create crime-identifying 'LawBot' », (17 octobre 2016), en ligne: *Legal Cheek* <https://www.legalcheek.com/2016/10/cambridge-university-law-students-create-crime-identifying-lawbot/>.

Corriveau, Noel, « Regulating automated decision systems in Canada: What it means for your business », (23 décembre 2020), en ligne: *INQ Law* <https://inqdatalaw.medium.com/regulating-automated-decision-systems-in-canada-what-it-means-for-your-business-bdbb04d6c725>.

Cybergenetics, « Cybergenetics complements the crime laboratory, making more science happen », (2021), en ligne: *Cybergenetics* <https://www.cybgen.com/services/crime_lab_complementor/page.shtml>.

Dastin, Jeffrey, « Amazon scraps secret AI recruiting tool that showed bias against women », *Reuters* (10 octobre 2018), en ligne: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

Data Skeptic, « Adversarial Explanations », (2020), en ligne: *Youtube* <https://www.youtube.com/watch?v=L4HCAow3W3Y>.

Dickson, Ben, « Self-supervised learning: The plan to make deep learning data-efficient », (23 mars 2020), en ligne: *TechTalks* <https://bdtechtalks.com/2020/03/23/yann-lecun-self-supervised-learning/>.

———, « The Book of Why: Exploring the missing piece of artificial intelligence », (9 décembre 2019), en ligne: *TechTalks* <https://bdtechtalks.com/2019/12/09/judea-pearl-the-book-of-why-ai-causality/>.

Domingo, Pedro, « The Five Tribes of Machine Learning », (2015), en ligne: *Association for Computing Machinery (ACM)* <https://learning.acm.org/binaries/content/assets/leaning-center/webinar-slides/2015/five-tribes-ml_112415.pdf>.

EDUCBA, « Machine Learning Tutorial | Self Guides to Learn Machine Learning », (2020), en ligne: *EDUCBA* <https://www.educba.com/data-science/data-science-tutorials/machine-learning-tutorial/>.

Equivant, « Northpointe Suite Risk Need Assessments », (mai 2021), en ligne: *Equivant* <www.equivant.com/northpointe-risk-need-assessments/>.

Face^{++}, « Face^{++} Cognitive Services », (2021), en ligne: *Face^{++}* <https://www.faceplusplus.com/>.

Félicien Vallet, « Jean-François Bonastre : “La voix n’est pas une biométrie classique” », (février 2017), en ligne: *Laboratoire d’innovation numérique de la CNIL* <https://linc.cnil.fr/fr/jean-francois-bonastre-la-voix-nest-pas-une-biometrie-classique>.

Fenjiro, Youssef, « Machine learning for Banking: Loan approval use case », (7 septembre 2018), en ligne: *Medium* <https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3>.

France Inter, « Épisode 6: Chacun sa voix du 01 août 2015 », (août 2015), en ligne: *France Inter* <https://www.franceinter.fr/emissions/scenes-de-crime/scenes-de-crime-01-aout-2015>.

Fridman, Lex, « Deep Learning Basics: Introduction and Overview (MIT course 6.S094) », (2019), en ligne: *Youtube* <https://www.youtube.com/watch?v=O5xeyoRL95U&list=PLrAXtmErZgOeiKm4sgNOknGvNjby9efdf>.

Goldman, Sally A, « Computational learning theory », (1991), en ligne: <https://dl.acm.org/doi/pdf/10.5555/1882757.1882783>.

Gooding, Matt, « “Robo-Lawyer” Can Predict How Likely You Are to Win a Case », (20 juin 2017), en ligne: *CambridgeshireLive* <www.cambridge-news.co.uk/business/technology/cambridges-lawbot-robo-lawyer-can-13209877>.

Google, « Assistant Google – Votre Google personnel », (novembre 2021), en ligne: *Google* <https://assistant.google.com/intl/fr_ca/>.

———, « Introduction to Machine Learning Problem Framing - Common ML Problems », (5 février 2021), en ligne: *Google Developers* <https://developers.google.com/machine-learning/problem-framing/cases?hl=fr>.

———, « What-If Tool », (novembre 2021), en ligne: *Google* <https://pair-code.github.io/what-if-tool/>.

Google Cloud Tech, « Using the What-If Tool for explainability », (2020), en ligne: *Youtube* <https://www.youtube.com/watch?v=jHojeFCc5HE>.

Grimson, Eric, John Guttag et Ana Bell, « Introduction to Computational Thinking and Data Science - MIT Course Number 6.0002 », (Automne 2016), en ligne: *MIT OpenCourseWare* <https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/>.

Guillaume Laberge et Éric Lavallée, « L'intelligence artificielle, bientôt réglementée au Canada ? », (29 janvier 2021), en ligne: *Lavery* <https://edoctrine.caij.qc.ca/publications-cabinets/lavery/2021/a121811/fr/i3fbc498f-432f-46da-8675-a5e7ec521452>.

Heaven, Will Douglas, « AI is wrestling with a replication crisis », (12 novembre 2020), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/>.

Hill, Kashmir, « Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match », *The New York Times* (29 décembre 2020), en ligne: <https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html>.

Hill, Kashmir, « The Secretive Company That Might End Privacy as We Know It », *The New York Times* (18 janvier 2020), en ligne: <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.

Hill, Kashmir, « Wrongfully Accused by an Algorithm », *The New York Times* (24 juin 2020), en ligne: <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>.

IBM, « IBM Watson products », (21 juin 2021), en ligne: *IBM* <https://www.ibm.com/watson/products-services>.

ISO, « ISO/IEC TR 24027:2021 Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making », (novembre 2021), en ligne: <https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/76/77607.html>.

Jacob Humerick, « Reprogramming Fairness: Affirmative Action in Algorithmic Criminal Sentencing », (15 avril 2020), en ligne: *Colum HRLR online* <hrlr.law.columbia.edu/hrlr-online/reprogramming-fairness-affirmative-action-in-algorithmic-criminal-sentencing/#post-1397-_Toc37843186>.

Jee, Charlotte, « A biased medical algorithm favored white people for health-care programs », (25 octobre 2021), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/>.

Karen Hao et Jonathan Stray, « Can you make AI fairer than a judge? Play our courtroom algorithm game », (17 octobre 2019), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>.

Kasture, Niwratti, « 10 Essential Ways to Evaluate Machine Learning Model Performance », (31 octobre 2020), en ligne: *Medium* <https://medium.com/analytics-vidhya/10-essential-ways-to-evaluate-machine-learning-model-performance-6bf6e11f9502>.

Kavlakoglu, Eda, « AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference? », (27 juillet 2020), en ligne: *IBM* <https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks>.

Knight, Will, « The Dark Secret at the Heart of AI », (2017), en ligne: *MIT Technology Review* <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>.

Lavallée, Éric, « La définition juridique de l'intelligence artificielle évolue : différents pays, différentes approches », (mars 2020), en ligne: *Lavery* <https://edoctrine.caij.qc.ca/publications-cabinets/lavery/2020/a121811/fr/i73b2616b-c0b5-4351-954c-4cce21486918>.

Marcus, Gary, « Deep Learning: A Critical Appraisal » (2018) arXiv:180100631 [cs, stat], en ligne: <http://arxiv.org/abs/1801.00631>.

Matthias Spielkamp, « Inspecting Algorithms for Bias », (12 juin 2017), en ligne: *MIT Technology Review*

<https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/>.

Metwalli, Sara A, « 5 Tools to Detect and Eliminate Bias in Your Machine Learning Models », (2 mars 2021), en ligne: *Medium* <https://towardsdatascience.com/5-tools-to-detect-and-eliminate-bias-in-your-machine-learning-models-fb6c7b28b4f1>.

Microsoft, « Azure Cognitive Services », (2021), en ligne: *Microsoft* <https://azure.microsoft.com/fr-fr/services/cognitive-services/>.

Minitab Statistical Software, « Diagrammes de relations linéaires, non linéaires et monotones », (2019), en ligne: *Minitab* <https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/basics/linear-nonlinear-and-monotonic-relationships/>.

Molnar, Christoph, « Interpretable Machine Learning: A Guide for Making Black Box Models Explainable », (2021), en ligne: <https://christophm.github.io/interpretable-ml-book/index.html>.

Mordor Intelligence, « AI Software Market in Legal Industry - Growth, Trends, Covid-19 Impact, and Forecasts (2021 - 2026) », (2020), en ligne: <https://www.mordorintelligence.com/industry-reports/ai-software-market-in-legal-industry>.

Moreno, Alice, « Police scientifique : comment les voix et les sons sont analysés pour résoudre des enquêtes », (28 octobre 2020), en ligne: *RTL* <https://www.rtl.fr/actu/justice-faits-divers/police-scientifique-comment-les-voix-et-les-sons-sont-analyses-pour-resoudre-des-enquetes-7800912370>.

Mousmoulas, Christos, « Probability vs Statistics », (16 décembre 2019), en ligne: *Medium* <https://towardsdatascience.com/probability-vs-statistics-95f221cc74f7>.

Muldoon, Matt, « L'intelligence artificielle vocale : qu'est-ce que c'est ? », (24 mars 2021), en ligne: *ReadSpeaker AI* <https://www.readspeaker.ai/fr/blog/what-is-multilingual-neural-text-to-speech-and-how-can-it-help-your-business-12-copy-copy/>.

Murray, David, « Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases », (20 mars 2015), en ligne: *The Courier Mail* <https://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-

miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b>.

Nuance Communications, « Biometric Authentication | Strong Customer Authentication », (2021), en ligne: *Nuance Communications* <https://www.nuance.com/omni-channel-customer-engagement/authentication-and-fraud-prevention/biometric-authentication.html>.

———, « Dragon Speech Recognition - Get More Done by Voice | Nuance », (2021), en ligne: *Nuance Communications* <https://www.nuance.com/dragon.html>.

———, « Nuance Forensics Datasheet », (2018), en ligne: *Nuance Communications* <https://www.nuance.com/content/dam/nuance/en_us/collateral/enterprise/data-sheet/ds-nuance-forensics-en-us.pdf>.

Nunes, Capt Jason, « DOD Policy Ignores Machine Learning », (16 mars 2020), en ligne: *Signal Magazine* <https://www.afcea.org/content/dod-policy-ignores-machine-learning>.

Ouest-France avec AFP, « "On l'entend mourir" : le terrible appel aux secours d'Élodie Kulik diffusé au procès de Willy Bardon », (27 novembre 2019), en ligne: <https://www.ouest-france.fr/societe/justice/l-entend-mourir-le-terrible-appel-aux-secours-d-elodie-kulik-diffuse-au-proces-de-willy-bardon-6627833>.

Parcoursup, « Parcoursup: Entrez dans l'enseignement supérieur », (2021), en ligne: *Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation* <https://parcoursup.fr/>.

Patenaude, Pierre, *De l'expertise judiciaire dans le cadre du procès criminel et de la recherche de la vérité: quelques réflexions.*, Université de Sherbrooke, 1996.

Rackspace Technology, « AI and Machine Learning Research Report », (26 janvier 2021), en ligne: <https://www.rackspace.com/solve/succeeding-ai-ml>.

———, « New Global Rackspace Technology Study Uncovers Widespread Artificial Intelligence and Machine Learning Knowledge Gap », (3 février 2021), en ligne: *Rackspace Technology* <https://www.rackspace.com/newsroom/new-global-rackspace-technology-study-uncovers-widespread-artificial-intelligence-and-0>.

Royaume-Uni, Centre for Data Ethics and Innovation (CDEI), « Facial Recognition Technology - Snapshot Paper », (28 mai 2020), en ligne: <https://www.gov.uk/government/publications/cdei-publishes-briefing-paper-on-facial-recognition-technology/snapshot-paper-facial-recognition-technology>.

Sarkar, Tirthajyoti, « Google's New "Explainable AI" (xAI) Service », (2 janvier 2020), en ligne: *Medium* <https://towardsdatascience.com/googles-new-explainable-ai-xai-service-83a7bc823773>.

Scherman, Michael et al, « US Lawmakers propose Algorithmic Accountability Act intended to regulate AI, April 22, 2019, dans CyberLex : insights on cybersecurity, privacy and data protection law, McCarthy Tétrault », (avril 2019), en ligne: <https://edoctrine.caij.qc.ca/publications-cabinets/mccarthy/2019/a98358/en/PC-a115788>.

Schmelzer, Ron, « Understanding Explainable AI », (23 juillet 2019), en ligne: *Forbes* <https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/>.

Scientific Analytical Tools, « BATVOX », (2017), en ligne: *Scientific Analytical Tools* <https://sat.ae/audio-forensics/forensic-analysis-tools/batvox/>.

Sermondadaz, Sarah, « L'algorithme de Parcoursup décrypté par les deux chercheurs qui l'ont conçu », (12 juin 2018), en ligne: *Sciences et Avenir* <https://www.sciencesetavenir.fr/high-tech/informatique/bac-2018-l-algorithme-de-parcoursup-explique-par-les-deux-chercheurs-qui-l-ont-concu_124407>.

Steven Pinker, « GENED 1066: Rationality », (2019), en ligne: *Harvard* <https://harvard.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx#folderID=%2255a37adc-eaae-4aa6-8a06-ab25015a4ee8%22&page=0>.

Stewart, Matthew, « The Limitations of Machine Learning », (29 juillet 2019), en ligne: *Towards Data Science* <https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6>.

Studycom, « Discontinuous Functions: Properties & Examples - Video & Lesson Transcript », (novembre 2021), en ligne: *Study.com* <https://study.com/academy/lesson/discontinuous-functions-properties-examples-quiz.html>.

ULCC, « Conférence pour l'harmonisation des lois au Canada », en ligne: <https://www.ulcc-chlc.ca/?lang=fr-ca>.

University of Oxford, « AI modelling tool developed by Oxford academics incorporated into Amazon anti-bias software », (avril 2021), en ligne: *Oxford Internet Institute* <https://www.oii.ox.ac.uk/news/releases/ai-modelling-tool-developed-by-oxford-academics-incorporated-into-amazon-anti-bias-software-2/>.

Vanessa, Henri, Deneault-Rouillard William et Agaby Linda, « Projet de loi n° 64 : Nouvelles règles encadrant la prise de décision individuelle automatisée. », (octobre 2020), en ligne: *Fasken* <https://edoctrine.caij.qc.ca/publications-cabinets/fasken/2020/a121765/fr/iac9cf9f8-7d17-47b1-9ff4-c61a33ca0af0>.

Villa, Jennifer et Yoav Zimmerman, « Reproducibility in ML: why it matters and how to achieve it », (25 mai 2018), en ligne: *Determined AI* <https://determined.ai/blog/reproducibility-in-ml/>.

Vincent, James, « The Biggest Headache in Machine Learning? Cleaning Dirty Data Off the Spreadsheets », (1 novembre 2017), en ligne: *The Verge* <https://www.theverge.com/2017/11/1/16589246/machine-learning-data-science-dirty-data-kaggle-survey-2017>.

Walch, Kathleen, « AI Laws Are Coming », (20 février 2020), en ligne: *Forbes* <https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/>.

Zorio, Stephen, « How a paper by three Oxford academics influenced AWS bias and explainability software », (1 avril 2021), en ligne: *Amazon Science* <https://www.amazon.science/latest-news/how-a-paper-by-three-oxford-academics-influenced-aws-bias-and-explainability-software>.

« IEEE 7000 Projects: IEEE Ethics In Action in Autonomous and Intelligent Systems », (2021), en ligne: *IEEE* <http://ethicsinaction.ieee.org/p7000/>