Évaluation structurée des besoins, exigences et contraintes en intelligence artificielle explicable pour une intelligence artificielle de confiance

par

Camélia RAYMOND

MÉMOIRE PAR ARTICLES PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLEÀ L'OBTENTION DE M. SC. A. EN GÉNIE LOGICIEL AVEC MÉMOIRE

MONTRÉAL, LE 15 MAI 2026

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

Mme Sylvie Ratté, directeur de thèse
Département de génie logiciel et TI à l'École de technologie supérieure

M. Marc-Kevin Daoust, codirecteur de thèse
Département des enseignements généraux à l'École de technologie supérieure

M. Mathias Glaus, président du jury
Département de génie de la construction à l'École de technologie supérieure

M. Luc Duong, membre du jury
Département de génie logiciel et TI à l'École de technologie supérieure

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 15 AVRIL 2026

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Ces trois dernières années, la réalisation de cette recherche a occupé mon esprit, mon temps, et surtout mes discussions. Elle a été à la fois ma force, et par moments, mon plus grand défi. Heureusement, tous ont su m'aider, à leur façon, à réaliser ce projet d'envergure, à aller au bout de ma passion. Mes remerciements les plus sincères vont à ces personnes qui m'ont permis d'avancer.

Mes sincères remerciements s'adressent à mes directeurs de recherche, Sylvie Ratté et Marc-Kevin Daoust, dont le soutien et les conseils ont été des piliers fondamentaux dans mon parcours académique. Leur expertise et leur approche méthodique m'ont non seulement guidé dans la complexité de la recherche, mais ont également façonné ma compréhension et mon appréciation de la rigueur scientifique. Leur capacité à éclairer les zones d'ombre, à orienter mes questionnements et à affiner mes hypothèses a transformé ce voyage intellectuel en une expérience enrichissante et formatrice. Sans leur précieuse collaboration, l'apport de mes recherches dans ce domaine n'aurait pas été aussi significatif.

Je tiens également à exprimer ma gratitude envers mon employeur, le Centre de Recherche Informatique de Montréal (CRIM), pour avoir cultivé un environnement de travail où la quête du savoir est grandement valorisée. Cette culture a facilité mon parcours académique. Un remerciement particulier va à Michel Savard et Martin Sortir, deux collègues exceptionnels, qui ont toujours été disponibles pour des échanges fructueux et des discussions stimulantes. Leur aide a été déterminante pour surmonter les obstacles rencontrés lors de la rédaction de mon mémoire. Enfin, je voudrais adresser une reconnaissance spéciale à Jordan Gieschendorf, ami et collègue de travail, pour son soutien constant et ses encouragements, qui ont joué un rôle crucial dans la réalisation de mes ambitions. Ambitions mêlant mes objectifs professionnels et académiques.

Je souhaite adresser mes remerciements les plus chaleureux à Cynthia Chassigneux, une figure éminente et respectée dans le secteur juridique québécois. Sa générosité tant du point de vue du temps passé, du partage des connaissances et de son soutien a été inestimable pour moi.

Un remerciement tout particulier est adressé à Hakima Hamdi, dont le rôle dans la correction de ma recherche a été indéniable. En tant que personne confrontée aux défis de la dyslexie et de la dysorthographie, l'apport d'Hakima a été pour moi d'une valeur inestimable. Sa capacité à préciser mes pensées écrites a grandement soutenu la qualité de mon travail.

Je tiens à souligner l'immense support de mes parents. Ma mère, en particulier, a été une source constante d'encouragement, me poussant à atteindre et même dépasser mes objectifs académiques. Son propre parcours, ayant réussi à obtenir une maîtrise tout en travaillant à temps plein, et enlevant deux petites filles, a été une source d'inspiration pour moi. Son exemple m'a donné la confiance et la force de croire en mes propres capacités. Mon père, quant à lui, a été l'architecte de mon parcours académique. Il m'a non seulement transmis ses connaissances, mais aussi son éthique de travail, sa passion pour l'excellence, son amour pour l'informatique et les mathématiques. Leur soutien indéfectible a été un facteur clé dans la réalisation de ce travail.

Ces trois années, entre le tourment et le rêve, ont représenté mon projet le plus ambitieux à ce jour. Merci Arthur de m'avoir soutenue dans ma passion. En cette dernière année particulièrement mouvementée, tu m'as entourée de ton amour et de ta générosité, me permettant de traverser les défis avec plus de sérénité. Ce projet maintenant terminé, je me réjouis à l'idée de me consacrer à nos projets communs, fort de cette expérience partagée.

Camélia

Évaluation structurée des besoins, exigences et contraintes en intelligence artificielle explicable pour une intelligence artificielle de confiance

Camélia RAYMOND

RÉSUMÉ

Au Québec, le déploiement de l'intelligence artificielle (IA) soulève des enjeux croissants de transparence, d'éthique et de responsabilité. Bien que la Loi 25 encadre désormais la gouvernance des systèmes automatisés, la mise en œuvre de l'intelligence artificielle explicable (XIA) demeure difficile à concrétiser dans les milieux industriels. Les approches actuelles se concentrent sur les aspects techniques des algorithmes, sans offrir de méthodologie claire pour cerner et formaliser les besoins réels des parties prenantes. Cette recherche vise à combler ce décalage en proposant une méthodologie pour identifier, structurer et prioriser les exigences, contraintes et besoins en XIA, adaptée au contexte québécois et alignée sur les principes de l'IA de confiance. Elle s'articule autour de quatre articles complémentaires explorant les dimensions juridiques, éthiques, organisationnelles et pratique de la XIA. Les trois premiers articles jettent les bases conceptuelles et normatives nécessaires à la conception d'une IA explicable : interprétation du cadre juridique québécois, opérationnalisation des principes éthiques et caractérisation des parties prenantes. Le quatrième article consolide ces acquis dans une méthodologie complète, le XAIRS (Explainable Artificial Intelligence Requirements Specification), un outil destiné à guider la spécification des exigences en XIA tout au long du cycle de vie des systèmes d'IA. Cette recherche met en évidence que la XIA peut être abordée comme un enjeu d'ingénierie, où la formalisation des besoins et des contraintes devient un levier de transparence et de redevabilité dans la conception des systèmes d'IA. Les résultats démontrent la faisabilité et la pertinence du XAIRS dans les contextes industriels québécois.

Mots-clés : Intelligence artificielle, IA de confiance, IA eXplicable, IA légale, IA éthique

Structured assessment of needs, requirements, and constraints in explainable artificial intelligence for trustworthy artifical intelligence

Camélia RAYMOND

ABSTRACT

In Quebec, the deployment of artificial intelligence (AI) raises growing issues of transparency, ethics, and accountability. Although Bill 25 now regulates the governance of automated systems, the implementation of explainable artificial intelligence (XIA) remains difficult to achieve in industrial settings. Current approaches focus on the technical aspects of algorithms, without offering a clear methodology for identifying and formalizing the real needs of stakeholders. This research aims to bridge this gap by proposing a methodology for identifying, structuring, and prioritizing XAI requirements, constraints, and needs, adapted to the Quebec context and aligned with the principles of trustworthy AI. It is structured around four complementary articles exploring the legal, ethical, organizational, and practical dimensions of XAI. The first three articles lay the conceptual and normative foundations necessary for the design of explainable AI: interpretation of the Quebec legal framework, operationalization of ethical principles, and characterization of stakeholders. The fourth article consolidates these findings into a comprehensive methodology, XAIRS (Explainable Artificial Intelligence Requirements Specification), a tool designed to guide the specification of XAI requirements throughout the life cycle of AI systems. This research highlights that XAI can be approached as an engineering issue, where the formalization of needs and constraints becomes a lever for transparency and accountability in the design of AI systems. The results demonstrate the feasibility and relevance of XAIRS in Quebec industrial contexts.

Key words : Artificial Intelligence, Thrustworthy AI, eXplainable AI, Legal AI, Ethical AI

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AI Act	Règlement européen sur l'intelligence artificielle
APEC	Association des praticiens en éthique du Canada
AUC	Area Under the Curve (Aire sous la courbe)
CEIMIA	Centre d'Expertise International de Montréal en IA
CEST	Commission de l'éthique en science et en technologie
CNRC	Conseil national de recherches du Canada
CRIM	Centre de recherche informatique de Montréal
CRISP-DM	Cross Industry Standard Process for Data Mining
DCAI	Data-Centric AI (IA centrée sur les données)
DDIRIA	Déclaration de Montréal pour un développement responsable de l'IA
ETS / ÉTS	École de technologie supérieure
IA	Intelligence artificielle
IEEE	Institute of Electrical and Electronics Engineers
ISO/IEC	Organisation internationale de normalisation / Commission électrotechnique internationale
LIAD	Loi sur l'intelligence artificielle et les données (projet de loi C-27)
LIME	Local Interpretable Model-Agnostic Explanations
LPRPSP	Loi sur la protection des renseignements personnels dans le secteur privé (Loi 25)
LPVPC	Loi sur la protection de la vie privée des consommateurs (projet de loi C-27)
MCAI	Model-Centric AI (IA centrée sur le modèle)
OSFI	Office of the Superintendent of Financial Institutions (Canada)
PME	Petite et moyenne entreprise

SHAP	SHapley Additive exPlanations
SIA	Système d'intelligence artificielle
UE	Union européenne
XAIRS	Explainable Artificial Intelligence Requirements Specification

INTRODUCTION

Au Québec, l'intelligence artificielle (IA) suscite une multitude de réactions parmi la population. Tandis que certains citoyens peinent à saisir l'IA, d'autres cherchent activement à la comprendre. L'IA s'intègre de plus en plus à divers secteurs de notre société. Cette progression a toutefois fait naître des préoccupations quant à la transparence, l'intégrité et l'éthique entourant son utilisation.

Dans les dernières années, l'IA a progressivement quitté le cadre expérimental pour s'intégrer à des environnements décisionnels ayant des conséquences directes sur la population. Des systèmes algorithmiques sont mobilisés dans des domaines tels que la santé, l'éducation, les ressources humaines, les services publics, la finance et la sécurité. Cette transformation modifie la nature du rapport entre la population et les technologies numériques. L'IA n'est plus uniquement un outil d'automatisation. C'est un intermédiaire capable d'influencer l'accès à des services, la priorisation de certaines personnes, l'évaluation de risques ou la recommandation d'actions. Les enjeux soulever par l'IA ne peuvent plus être réduits à des questions de performance techniques. Ils concernent aussi la justification des décisions, la distribution des bénéfices et des risques et la capacité des personnes affectées à comprendre, questionner ou contester les résultats produits.

Cette préoccupation est documentée par plusieurs travaux ayant montré que des systèmes d'IA peuvent reproduire, amplifier ou rendre moins visibles certaines inégalités sociales. Par exemple, en santé mentale, Pichowicz, Kotas et Piotrowski (2025) ont évalué 29 agents conversationnels à l'aide de scénarios simulant une crise suicidaire. Plus de la moitié des systèmes n'ont pas fourni d'information d'urgence permettant de rejoindre une aide humaine, alors qu'il s'agit d'un critère minimal pour répondre à une situation de risque suicidaire. Dans le domaine de la reconnaissance faciale, l'erreur peut prendre une forme particulièrement grave. L'on peut associer le visage d'une personne innocente à celui d'un suspect à partir d'une image de surveillance imparfaite. Les évaluations du National Institute of Standards and Technology montrent que ce type d'erreur peut varier selon certains groupes démographiques

et, en contexte policier, mener à de fausses accusations (NIST, 2019; NIST, 2020). Ce risque n'est pas seulement théorique : une enquête du Washington Post a recensé au moins huit arrestations injustifiées liées à la reconnaissance faciale, dont celle d'un homme détenu pendant plus de seize mois après avoir été associé à une image granuleuse par un système automatisé (Washington Post, 2025).

Dans ce contexte, l'« IA de confiance » (terminologie définie dans le chapitre 1) a été proposée comme étant une solution nécessaire à la création et à l'implémentation d'IA répondant aux craintes du public. Cette implémentation est basée sur plusieurs caractéristiques, l'une d'elle étant la transparence des IA. Cela implique que les parties prenantes d'un système d'IA (SIA) comprennent la logique sous-jacente de l'algorithme. Bien que des techniques d'explicabilité soient disponibles afin de répondre à cette demande, celles-ci semblent insuffisantes pour répondre aux besoins spécifiques de l'industrie. Nous observons un manque criant de procédures claires pour définir les besoins en IA eXplicable (XIA). Cette lacune rend complexe le développement de techniques d'explicabilité adaptées aux attentes des consommateurs.

La présente recherche, qui adopte une démarche scientifique et basée sur les principes d'ingénierie, se propose de combler le décalage persistant entre les explications générées par les techniques de la XIA et les besoins réels des parties prenantes, en vue d'améliorer la légitimité de ces systèmes dans l'industrie québécoise. Nous souhaitons proposer une méthodologie qui permet de cerner et définir les besoins, exigences et contraintes de la XIA. En focalisant sur le paysage québécois, cette méthodologie augmente significativement les chances que la XIA soit non seulement en accord avec les réglementations en vigueur (licite) mais aussi conforme aux principes éthiques.

Nous cherchons à offrir une contribution multidimensionnelle à la discipline de l'IA, spécifiquement dans le domaine de la XIA. Voici les principales contributions visées.

1) Protocole de définition des parties prenantes : Introduire un protocole qui identifie et catégorise les différentes parties prenantes en relation avec un système de XIA. Cette

démarche répond à un manque notable dans la littérature et constitue une avancée significative pour cerner les besoins en XIA.

2) Définition des connaissances des parties prenantes : Proposer une méthodologie structurée pour évaluer les connaissances applicables au domaine de l'IA des parties prenantes. Cela permet de comprendre les besoins en matière de vulgarisation, justifiant ainsi la pertinence et l'efficacité des explications fournies par les techniques de XIA.
3) Protocole de définition des besoins, exigences et contraintes en XIA : Fournir un cadre de référence pour la compréhension des besoins en XIA, en intégrant des perspectives légales et éthiques, souvent négligées ou peu investiguées dans les études antérieures. Celui-ci sera bâti sur la base des concepts développés dans le protocole de définition des parties prenantes ainsi que l'évaluation de leurs connaissances.
4) Méthodologie de priorisation : Apporter une approche structurée pour hiérarchiser les besoins en XIA, comblant ainsi une lacune dans les démarches actuelles. Cette méthodologie assure que les efforts de développement et d'implémentation soient dirigés vers les aspects les plus critiques.
5) Interprétation juridique appliquée à la XIA : Combler le fossé entre la loi et la pratique en identifiant les écarts existants entre les exigences juridiques récentes au Québec et les réalités du développement des SIA explicables. Cette contribution propose une interprétation concrète des obligations légales liées à la XIA, en les traduisant en recommandations pratiques adaptées au contexte de l'ingénierie et de la gestion de projets en IA.
6) Opérationnalisation des principes éthiques : Fournir aux ingénieurs une approche structurée pour traduire les principes éthiques en mesures concrètes applicables dans le contexte de la XIA.

Chaque contribution développée dans le cadre de cette recherche a été intégrée et testée dans une dernière contribution synthèse, regroupant l'ensemble des éléments proposés au fil du travail (présentée au chapitre 6). Chaque contribution est soumise à une évaluation quantitative au moyen d'une enquête réalisée auprès du responsable technique d'un projet de recherche industriel dans un secteur critique du domaine de l'intelligence artificielle qui nécessite de la XIA. L'objectif de ce sondage est de mesurer l'utilité et l'efficacité des apports fournis par cette thèse.

Pour ce faire, nous explorons, au chapitre 1, les enjeux transversaux qui entourent l'utilisation de la XIA. Dans un premier temps, nous plaçons l'importance de l'IA de confiance dans le contexte québécois et définissons la contribution de la XIA. Ensuite, nous explorons les contraintes de la XIA en validant les lois québécoises et canadiennes qui régissent son utilisation. Les principes éthiques et les mesures concrètes pour les respecter ont fait l'objet d'une investigation approfondie. Les enjeux relatifs à la robustesse des SIA et l'apport de la

XIA pour contrer ceux-ci sont mis en lumière dans les cas d'usage. Cela nous amène ensuite à définir les métriques d'évaluation des techniques de XIA et de mettre ainsi l'emphase sur l'absence de celles-ci. La revue de littérature est complétée par la définition du cycle de vie des SIA ainsi que l'analyse des parties prenantes de celui-ci.

Au chapitre 2, nous présentons les prochaines étapes qui nous permettrons de créer une méthodologie pour cerner et définir les besoins, exigences et contraintes de la XIA. Les chapitres 3 et 4 traduisent des exigences normatives et éthiques de la XIA en éléments directement mobilisables par les ingénieurs IA. Le chapitre 5 propose un protocole d'identification des parties prenantes reposant sur une double caractérisation : d'une part, leur rôle au sein du cycle de vie d'un système d'IA (utilisateur, développeur, décideur, régulateur, etc.), et d'autre part, leur niveau d'expertise en intelligence artificielle. Ce qui nous amène, au chapitre 6, à présenter une méthodologie structurée pour formaliser les besoins, exigences et contraintes en matière de la XIA. Finalement, le chapitre 7 discute des impacts de ses contributions de manière individuelles et combinée. Nous explorons les pistes possibles d'améliorations.

CHAPITRE 1

REVUE DE LITTÉRATURE

De plus en plus de secteurs sont concernés par l'IA. Ainsi, on rencontre ces algorithmes dits intelligents dans les industries, les banques, les hôpitaux, les cliniques ou les compagnies d'assurance. Cependant, dans certains de ces secteurs, les exigences de traçabilité et d'explicabilité des résultats sont plus critiques. Les utilisateurs ont besoin d'algorithmes explicables afin d'augmenter la confiance envers les résultats que ces algorithmes proposent ainsi que pour comprendre la provenance des résultats selon une perspective scientifique. Ce chapitre a pour objectif de développer ces deux aspects, explicabilité et confiance, afin d'avoir une compréhension juste et précise des problèmes que l'utilisation de l'IA explicable soulève.

Dans la première partie de ce chapitre, section 1.1, nous explorons la présence et le rôle des SIA au Québec, mettant en lumière leur impact significatif sur l'économie québécoise ainsi que dans le secteur académique. Cette analyse nous permet de souligner (section 1.2) l'importance de la confiance dans ces systèmes pour leur intégration réussie et leur adoption. Puis, le chapitre se penche sur le concept fondamental d'une IA de confiance, accessible, en partie, grâce à l'explicabilité des algorithmes d'IA. Nous établirons que pour qu'une IA explicable soit de confiance, elle doit répondre à plusieurs critères essentiels; elle doit être (1) licite (section 1.3), (2) éthique (section 1.4) et (3) robuste (section 1.5), et ce, à travers tout (4) son cycle de vie (section 1.6) et vis-à-vis de toutes (5) les parties prenantes (section 1.7). Ainsi, chacun de ses 5 thèmes est étudié. Cette approche nous offre un regard d'ensemble sur les concepts et met en lumière les lacunes existantes dans la littérature concernant les XIA.

1.1 Système d'intelligence artificielle au Québec

Plusieurs définitions ont été proposées pour définir l'IA. Nous retiendrons celle émise par le Gouvernement du Canada dans le Projet de Loi C-27 qui introduit la Loi sur l'intelligence artificielle et les données.

La Loi sur l'intelligence artificielle et les données (2023) définit une IA comme « Système technologique qui, de manière autonome ou partiellement autonome, traite des données liées à l'activité humaine par l'utilisation d'algorithmes génétiques, de réseaux neuronaux, d'apprentissage automatique ou d'autres techniques pour générer du contenu, faire des prédictions ou des recommandations ou prendre des décisions » (Art. 2).

Un système d'intelligence artificielle (SIA), est un assemblage de briques logicielles et matérielles permettant de réaliser une fonction ou un service, soit « purement » numérique, soit produisant un actionnement sur le monde physique (Chiaroni, 2021). Selon Chiaroni (2021), un système d'IA comprend trois composantes principales : des données, un algorithme ou une somme d'algorithmes ainsi qu'un composant ou une architecture électronique. Cette recherche sera focalisée sur la XIA au niveau des algorithmes.

La section 1.1.1 se penche sur le portrait de l'IA au Québec. À la section 1.1.2, nous mettons en évidence les distinctions entre le contexte industriel et le contexte académique du développement des algorithmes liés à l'IA. Il en ressort que l'opacité des algorithmes, entre autres, limite leur utilisation dans l'industrie. Ce phénomène d'opacité est appelé « boîte noire » (« black box »), dans le sens où l'industrie connaît les données qui entrent, constatent les résultats qui en sortent, mais ne comprennent pas comment on arrive à ces résultats.

1.1.1 Portrait de l'intelligence artificielle au Québec

En 2019, le Financial Times a évalué les 20 plus grandes villes d'Amérique du Nord en matière d'investissement en IA. La ville de Montréal s'est classée en première position dans ce palmarès grâce à la qualité et la compétitivité de ses investissements (Marcellis-Warin et al., 2021). Il n'est pas étonnant que les entreprises désirent s'installer à Montréal. En effet, le Québec est reconnu internationalement comme étant le pôle de l'IA, et par conséquent attire de plus en plus d'entreprises utilisant des SIA. DeepMind, Facebook, Google, Microsoft, Samsung, par exemple, ont choisi d'investir dans le secteur de l'IA à Montréal. La métropole se classe troisième à l'échelle internationale pour ce qui est du nombre de PME offrant de l'IA

dans une même ville (Marcellis-Warin et al., 2021). De plus, cette technologie est fortement encouragée et soutenue par le gouvernement du Québec qui octroie diverses subventions pour aider les entreprises (Gouvernement du Québec, 2021b).

D'un point de vue économique, plusieurs cabinets conseils s'avancent sur le potentiel de l'IA dans l'industrie. Par exemple, les analyses du cabinet de conseil Accenture estiment que l'IA pourrait doubler les taux de croissance économique annuels et augmenter jusqu'à 40% la productivité du travail dans les pays développés d'ici à 2035, ce qui permettrait d'accroître de près de 38% la rentabilité des entreprises (Marcellis-Warin et al., 2021). Plus précisément, l'IA au Québec a permis de créer des retombées économiques liées aux investissements de 237,8 millions au provincial et 155,7 millions au fédéral (Price Waterhouse Coopers, 2022). Avec tous ses chiffres faramineux, il n'est pas surprenant de compter au Québec plus de 600 organisations offrant des produits ou des services propulsés par l'IA (Price Waterhouse Coopers, 2022). Les entreprises québécoises reconnaissent le potentiel de l'IA et souhaitent l'utiliser au profit de leur développement économique.

Au Québec, le secteur de la recherche est un acteur très important en IA et joue un rôle majeur dans l'attrait des entreprises voulant faire de l'IA. Avec plus de 11 universités offrant des programmes liés à l'IA, le nombre d'inscriptions en mathématiques et informatique a crû de 16% en moyenne par an entre 2017 et 2020 (Price Waterhouse Coopers, 2022). Au Québec uniquement, dans les années 2020 et 2021, le nombre de chaires de recherche en IA titulaires de subventions provenant du gouvernement fédéral a augmenté de 400% (Price Waterhouse Coopers, 2022). Ces financements au niveau de la recherche ont comme objectif d'attirer les meilleurs talents en IA au Québec et de produire une main-d'œuvre de qualité pour les entreprises qui y sont installées.

1.1.2 Développement de l'intelligence artificielle dans un contexte académique en comparaison avec le contexte industriel

Si l'IA est désormais omniprésente dans les marchés scientifiques, économiques et politiques, son passage de la recherche à l'industrie ne va pas de soi. Le développement d'un modèle d'IA

en contexte académique diffère de sa mise en œuvre dans un environnement industriel. Ces différences, souvent sous-estimées, expliquent en partie pourquoi de nombreux prototypes prometteurs ne dépassent jamais le stade expérimental (Zaharia et al., 2018). L'un des principaux freins à ce transfert est le manque de transparence — ou « l'opacité » — de certains systèmes d'IA, ce qui entrave leur intégration dans des environnements régulés, éthiquement sensibles ou soumis à une gouvernance rigoureuse.

L'on parle souvent de « développement d'IA » comme si c'était un processus homogène, mais en réalité, les objectifs, contraintes et pratiques diffèrent significativement entre la recherche académique et l'industrie. En milieu académique, le développement d'un modèle d'IA vise souvent à maximiser la performance sur une tâche donnée, dans un cadre expérimental contrôlé. À l'inverse, en industrie, le modèle s'inscrit dans une chaîne de valeur, dans un environnement dynamique et fortement contraint par les exigences réglementaires, économiques et humaines. Les entreprises souhaitent résoudre des problèmes qui répondent à des enjeux d'affaires alors que les universités, par exemple, veulent résoudre des problèmes qui permettent de faire avancer les connaissances. Dans la pratique, cela se traduit par plusieurs différences notables, tel que le cycle de vie des modèles d'IA.

Dans les laboratoires de recherche, la conception d'un modèle d'IA suit un cycle relativement linéaire et axé sur la performance technique. Il s'agit souvent d'explorer un problème bien défini, de collecter des données (parfois synthétiques ou issues de jeux standards comme ImageNet ou MNIST), d'entraîner plusieurs modèles et de publier celui qui obtient les meilleurs résultats selon une métrique choisie (ex. : précision, F1-score, AUC). Les enjeux de maintenance, de déploiement ou de robustesse sont souvent secondaires, voire absents (Sculley et al., 2015).

En milieu industriel, le développement d'un système d'IA s'inscrit dans une chaîne de valeur complexe, avec des exigences métier, juridiques, technologiques et humaines. Le cadre méthodologique CRISP-DM (Cross Industry Standard Process for Data Mining) est encore largement utilisé pour structurer ces projets. Il comprend les étapes de compréhension du

contexte d'affaires, compréhension et préparation des données, modélisation, évaluation selon des critères métiers, déploiement opérationnel ainsi que la surveillance continue, l'audibilité et la gouvernance. Comme le souligne plusieurs entités qui définissent des normes en IA (Commission européenne, 2021 ; ISO/IEC JTC 1/SC 42, 2021), la robustesse, la résilience, la conformité réglementaire et la capacité à rendre compte des décisions automatisées sont des critères au moins aussi importants que la performance algorithmique en milieu industriel.

Ainsi, ce qui est tolérable (voire valorisé) dans un cadre académique – comme la recherche de la performance au détriment de la transparence – devient un obstacle dans un contexte industriel, où la fiabilité, la responsabilité et la capacité à rendre des comptes sont primordiales. Pour que les modèles d'IA passent à l'échelle dans l'industrie, ils doivent s'accompagner de garanties de transparence, d'équité, de sécurité et de contrôle humain. Bref, ils doivent être « de confiance ».

1.2 L'intelligence artificielle de confiance grâce à l'explicabilité

L'IA de confiance est un idéal difficile à définir. La section 1.2.1 permet de préciser ce concept en clarifiant les caractéristiques et les thèmes sous-jacents à l'IA de confiance. La sous-section suivante présente les SIA explicables comme une réponse au manque de confiance.

1.2.1 La théorie de la confiance

Lors d'une conférence donnée au Centre de recherche informatique de Montréal (CRIM), Hans Bherer[1] a proposé un parallèle intéressant entre le développement des ascenseurs et le développement des systèmes d'IA. En 1854, Elisha Graves Otis inventa et présenta au public un dispositif de sécurité qui empêche les ascenseurs de tomber en cas de défaillance du câble de levage. Avant sa démonstration devant le public, il n'existait aucun procédé permettant de sécuriser les monte-charges. Grâce à une démonstration à New-York, Otis démontra que son système de blocage en cas de rupture de corde était sûr et les conséquences en furent immenses.

[1] Directeur senior en recherche et technologie au Centre informatique de Montréal, docteur en philosophie, post doctorant en génie logiciel.

Dans les années qui suivirent, le marché immobilier a été complètement bouleversé. En effet, le mécanisme de verrouillage de sécurité ayant parfaitement fonctionné, les gens étaient de plus en plus disposés à monter dans des ascenseurs. Les ascenseurs ont contribué à rendre possible les gratte-ciels actuels. La confiance que le public avait acquise en cette technologie s'est avérée être très profitable pour Elisha Graves Otis et révolutionna plusieurs marchés. Au moins, l'une des conclusions à tirer de cette histoire est que, du point de vue commercial et marketing, l'opinion du public est cruciale.

Plusieurs définitions du terme « confiance » ont été utilisées dans le passé. L'une des plus récentes serait celle de Lee (2004) qui définit la confiance comme étant « un acte de foi » (leap of hope). L'espoir est l'attente d'un résultat favorable attendu par la personne qui fait confiance. Il convient de souligner que cette définition n'est pas universelle, différentes études utilisant des définitions autres que celle de Lee.

Une question fréquemment posée au sujet de la confiance est la suivante : Est-ce que la confiance est une caractéristique personnelle ou de groupe? Des chercheurs tels que Mayer et al. (1995) soutiennent la première hypothèse alors que Lewis et Weigert (1985) et Luhmann (1970) soutiennent la seconde. Par ailleurs, certaines notions telles que les types de confiance ont bien été établies et sont soutenues par une grande quantité de chercheurs. Il est possible de citer, entre autres, Lewis et Weigert (1995), Lewicki et al. (1998), Vanhala et al. (2016), qui soutiennent l'idée de deux types de confiance : cognitive et affective. La confiance cognitive se base sur la raison. Celle-ci s'appuie habituellement sur la preuve que quelque chose est assez bonne pour être digne de confiance. La confiance affective est émotionnelle et se fonde sur un sentiment de dévouement et de préoccupation envers la personne qui offre sa confiance. Selon Lewicki et al. (1998), la confiance cognitive apparaît généralement avant la confiance affective. En revanche, la confiance affective, une fois installée, peut difficilement être brisée contrairement à la confiance cognitive. Lewis et Weigert (1985) indiquent que ces deux formes de confiance coexistent dans une relation de confiance, mais possèdent des degrés d'importance différents en fonction de la personne qui accorde sa confiance.

Hofstede (1984) a présenté un modèle socioculturel qui permet de quantifier les niveaux de confiance d'une société dans différents domaines. Un des éléments de ce modèle concerne l'évitement de l'incertitude et se réfère à la (non) conformité du grand public vis-à-vis de la technologie, de la loi et de la religion (Hofstede, 1984). Le score, entre 1 et 100, est défini grâce à un sondage conçu pour évaluer les attitudes et les valeurs, y compris la tolérance à l'ambiguïté, les besoins de règles claires et la réaction face à des situations incertaines ou inconnues. Les cultures qui ont un indice élevé sont moins tolérantes face au changement. Présentement, le pays qui possède l'indice le plus élevé est la Grèce, avec un score de 100 (Hofstede Insights, 2023). Les sociétés dont l'indice est faible sont plus ouvertes au changement. La culture qui possède le score de 8, le plus petit du classement, est Singapore (Hofstede Insights, 2023).

Le Canada possède un score de 48, ce qui est considéré comme étant assez bas. Ce score indique une tolérance de la société pour l'incertitude. Cependant, les Canadiens francophones, qui se trouvent majoritairement au Québec (Statistics Canada, 2017), sont plus intolérants à l'incertitude que les Canadiens anglophones. Effectivement, on a constaté que les Québécois démontrent systématiquement des niveaux plus bas de confiance en général, que le reste du Canada (Brie, 2018). De plus, Denis et al. (2010) ont établi que la confiance des Canadiens francophones est associée à la capacité de bien performer une tâche, ce qui fait référence à la confiance cognitive. Nous émettons l'hypothèse, sans en connaître l'étendue, que la confiance cognitive est plus importante pour les Canadiens francophones que la confiance affective. Cette hypothèse devrait être prise en considération par les entreprises qui développent de nouvelles technologies, puisque tel que démontré avec l'anecdote de l'ascenseur, la confiance est un indicateur de succès.

En 2018, 35% des québécois se sont dit inquiets face au développement de l'IA (De Marcellis-Warin et al. 2021). Au Canada, ce sont près des deux tiers de la population qui partagent cette opinion (Innovation, Sciences et Développement économique Canada, 2021). Au Québec et au Canada, nous pouvons observer un manque de confiance envers l'IA. Plusieurs erreurs se sont produites dans le passé par des entreprises qui utilisaient des SIA. En 2015 par exemple,

Google a dû s'excuser publiquement parce que son algorithme de reconnaissance d'images à étiqueté des personnes Afro-américaines comme étant des « gorilles » (Zhang, 2015). En 2016, le logiciel Autopilote intégré à la voiture Tesla « Model S » a provoqué un accident mortel du conducteur. Il n'est pas clair que l'accident a été causé par un dysfonctionnement du système d'autopilotage et que l'algorithme a fonctionné conformément aux attentes et a donc pris la bonne décision (Knight, 2016). Ces incidents érodent bien évidemment la confiance que le public porte en l'IA. Afin de permettre aux entreprises d'atteindre le plein potentiel économique de l'IA, le cadre d'une utilisation d'IA digne de confiance permettant son acceptabilité doit devenir une priorité.

Même si plusieurs chercheurs, membres des gouvernements et organismes utilisent les termes « intelligence artificielle de confiance », peu de personnes semblent leur attribuer une définition. Middleton et al. (2022) définissent la confiance comme étant une attitude selon laquelle un agent se comportera comme prévu et sur lequel on peut compter pour atteindre son objectif. La confiance se brise après une erreur ou un malentendu entre l'agent et la personne qui fait confiance (Middleton et al., 2022). À la suite de ses recherches dans le domaine de l'éthique en l'IA, Ryan (2020) arrive à la conclusion suivante concernant la syntaxe d'une IA de confiance.

« On peut se fier à quelqu'un sur la base d'habitudes fiables, mais accorder sa confiance exige que l'autre agisse avec bonne volonté à l'égard de celui qui lui fait confiance. C'est la principale raison pour laquelle les objets fabriqués par l'homme, comme l'intelligence artificielle, peuvent être fiables, mais pas dignes de confiance » (Ryan, 2020).

En dépit de ses recherches et du haut niveau de reconnaissance scientifique reconnu, dans la mesure où les termes « intelligence artificielle de confiance » étaient déjà grandement utilisés dans l'industrie et dans les domaines académiques, l'expression « intelligence artificielle de confiance » a continué à être utilisée dans la littérature. Nous allons également l'utiliser pour nos travaux dans ce domaine. En 2019, la Commission Européenne et la Direction générale des réseaux de communication, du contenu et des technologies a publié Les lignes directrices

en matière d'éthique pour une IA digne de confiance (2019) qui définit l'IA de confiance comme suit :

« Une intelligence artificielle digne de confiance présente les trois caractéristiques suivantes, qui devraient être respectées tout au long du cycle de vie du système :
a) elle doit être licite, en assurant le respect des législations et réglementations applicables ;
b) elle doit être éthique, en assurant l'adhésion à des principes et valeurs éthiques ;
c) elle doit être robuste, autant sur le plan tant technique que social, car même avec de bonnes intentions, les systèmes d'intelligence artificielle peuvent causer des préjudices involontaires » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

Il est important de clarifier cette définition afin d'écarter les débats récents sur la définition du terme « confiance » et ainsi éclaircir sa signification dans l'expertise de l'IA. Les trois caractéristiques mentionnées ci-haut, soit être licite, éthique et robuste, sont considérées comme étant nécessaires à une IA digne de confiance, mais il n'est pas clair qu'elles soient suffisantes. En plus de l'énoncé de ces caractéristiques, cette définition introduit le fait que celles-ci doivent être respectées tout au long du cycle de vie d'un système d'IA.

1.2.2 Systèmes d'intelligence artificielle explicables en réponse au manque de confiance

Maintenant que nous avons établi les caractéristiques d'une IA de confiance, nous devons aussi introduire les principes clés d'une IA de confiance. Dans ce contexte, nous pouvons définir que les caractéristiques d'une IA de confiance désignent les qualités observables et mesurables qu'un SIA doit manifester tout au long de son cycle de vie pour être considéré comme digne de confiance. Elles représentent les résultats concrets que l'on cherche à atteindre, soit être licite, éthique et robuste, tout au long du cycle de vie. Les principes clés peuvent être définit comme des valeurs fondamentales et normatives qui guident la conception, le développement, le déploiement et la gouvernance des SIA. Ils forment un cadre stratégique à partir duquel les

caractéristiques souhaitées de l'IA peuvent être mises en œuvre. Les principes clés orientent l'action, tandis que les caractéristiques permettent d'en évaluer la réalisation.

Selon la Commission Européenne (2019). La confiance dans le domaine de l'IA s'obtient en implémentant sept principes clés : le contrôle humain, la robustesse et la sécurité, la conformité et la gouvernance, l'explicabilité et la traçabilité, la diversité et l'équité, l'impact social et environnemental, et les responsabilités associées au traitement. La figure 1.1 représente visuellement la définition des systèmes d'IA de confiance, démontrant les principes clés et les caractéristiques, selon la Commission Européenne (2019).

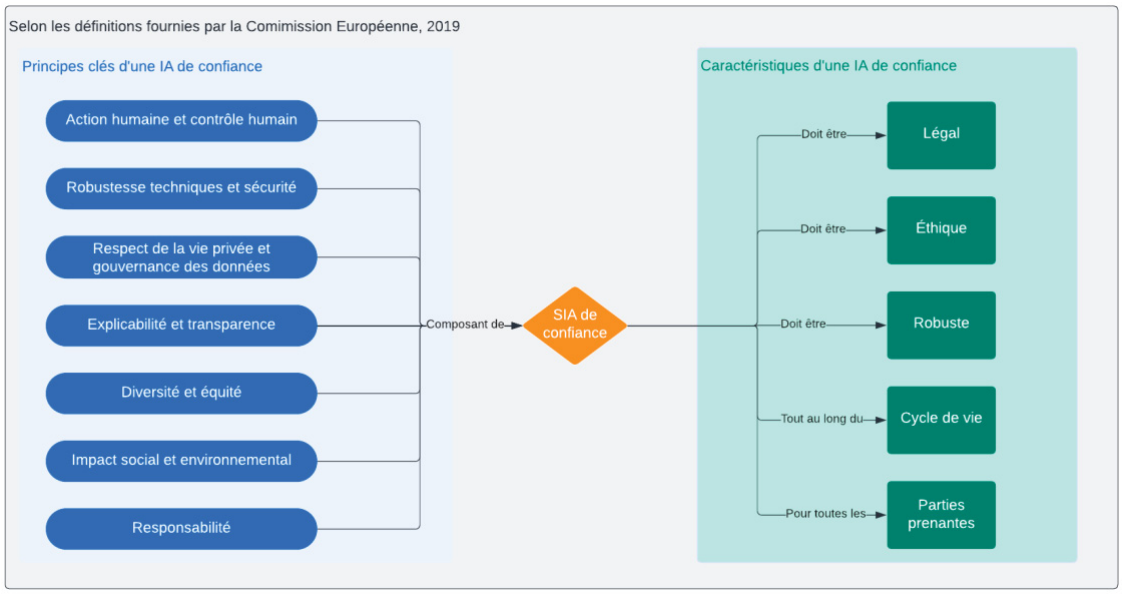


Figure 1.1 Représentation visuelle de la définition des systèmes d'IA de confiance, incluant les principes clés et les caractéristiques d'une IA de confiance

L'explicabilité des systèmes d'IA peut être perçue comme un morceau du casse-tête qu'est l'IA de confiance. Ce principe permet, conjointement avec d'autres principes, de répondre au manque de confiance en l'IA.

Un cas d'utilisation fréquemment décrit dans la littérature portant sur la XIA est l'analyse de décisions de prêt bancaire. Il s'agit d'une demande de financement personnel dans laquelle les clients demandent un prêt à une institution financière telle qu'une banque. Un SIA explicable, selon les suggestions de Arya *et al.* (2019), fournirait une liste des caractéristiques utilisée par l'IA et l'impact chiffré de ses caractéristiques sur la recommandation d'accepter ou de refuser le prêt. Arya *et al.* (2019) identifient jusqu'à cinq catégories de techniques relatives à la XIA, chaque catégorie comprenant jusqu'à dix méthodes distinctes pour leur mise en œuvre. Il n'existe pas de consensus sur une méthode particulière.

Selon les *Lignes directrices en matière d'éthique pour une IA digne de confiance*, le concept d'explicabilité se définit comme suit :

« L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'intelligence artificielle et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'intelligence artificielle). L'explicabilité technique suppose que les décisions prises par un système d'intelligence artificielle peuvent être comprises et retracées par des êtres humains. [...] Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur) » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

L'utilisation de SIA explicable permettrait de répondre, en grande partie, aux enjeux de transparence de la décision et ainsi renforcer et conserver la confiance des utilisateurs envers les systèmes d'IA. Cependant, la problématique de la boîte noire renvoie à l'incapacité à comprendre notamment les raisons ayant conduit l'IA à produire de tels résultats et rend difficile, voire impossible, l'explicabilité des décisions rendues par les SIA. Même les personnes qui développent l'IA ne sont pas toujours en mesure d'en expliquer la logique. Au Québec, un total de 44% des citoyens pensent savoir ce qu'est l'IA et 10% d'entre eux ne savent pas du tout ce que c'est (Marcellis-Warin et al. 2021). Seuls 13% affirment savoir précisément de quoi il s'agit (Marcellis-Warin et al. 2021). L'IA explicable répond

partiellement au manque de confiance envers cette technologie. Augmenter ce niveau de confiance de la clientèle permettrait aux entreprises de tirer profit de l'IA.

Afin de répondre aux exigences d'une IA de confiance, un système d'intelligence artificielle explicable doit, à tout le moins, présenter les mêmes caractéristiques fondamentales, à savoir : la licéité (section 1.3), l'éthique (section 1.4) et la robustesse (section 1.5). Conformément aux définitions proposées par l'Union européenne, l'explicabilité doit en outre être assurée tout au long du cycle de vie du système (section 1.6) et adaptée au niveau d'expertise des parties prenantes concernées (section 1.7). Cette revue de littérature se poursuit donc en examinant l'explicabilité comme un levier essentiel permettant d'atteindre les différentes caractéristiques d'une IA de confiance. Les sections suivantes sont chacune consacrées à l'une de ces caractéristiques, telles qu'illustrées à la figure 1.2.

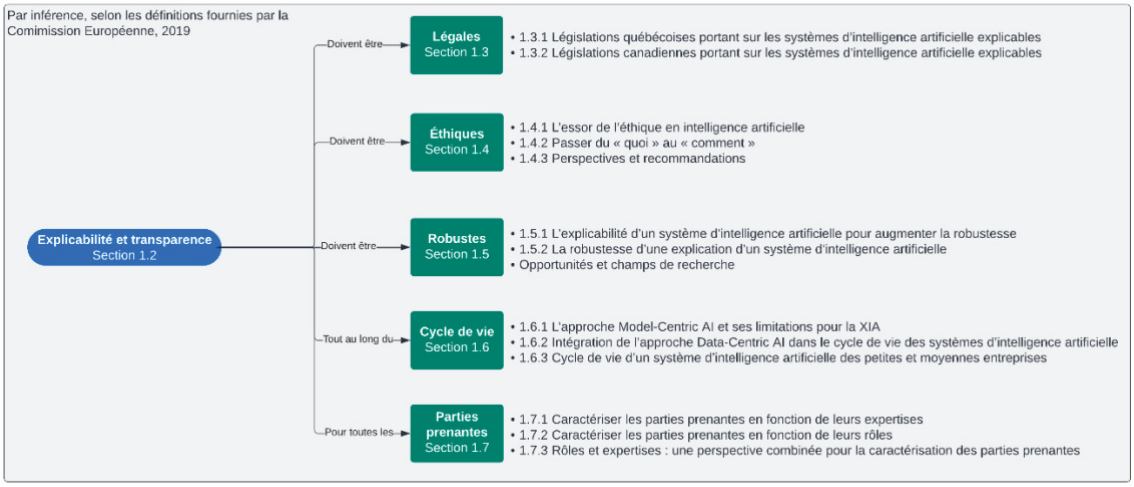


Figure 1.2 Représentation visuelle de la définition des systèmes d'IA de confiance, incluant les principes clés et les caractéristiques d'une IA de confiance

1.3 Les systèmes d'intelligence artificielle explicables et licites

L'explicabilité des systèmes d'IA n'est pas exclusivement une question commerciale qui permet de satisfaire les consommateurs. D'un point de vue juridique, l'explicabilité est un outil qui peut être utilisé pour effectuer une analyse post-incident révélant, autant que possible, ce

qui a conduit cet l'incident (Tobey, 2019). Alors que le Québec et le Canada ont décidé de prioriser le développement de nouvelles technologies ainsi que l'innovation (Assemblée nationale du Québec, 2019), l'Europe a misé sur la protection des consommateurs (Sartor et Lagioia, 2020) lors de l'élaboration des lois en matière d'IA explicable. L'utilisation de termes ambigus (par exemple : principaux facteurs et paramètres) dans la réglementation québécoise, plus particulièrement l'article 12.1 de la loi 64, laisse à penser que le Québec voudrait instaurer des obligations en termes d'utilisation d'IA explicable. Cependant, du fait de la nature encore expérimentale des recherches dans ce domaine, une législation trop contraignante pourrait limiter le développement et l'utilisation de l'IA au Québec. Au Canada, plusieurs commissions, rapports et recherches ont démontré la nécessité d'adapter les lois actuelles afin d'y inclure un droit à l'explication. Contrairement à son prédécesseur (Projet de Loi C-11), le nouveau Projet de Loi C-27 prend en considération ses recommandations de manière graduelle.

Les sections suivantes ont pour objectif de rapporter les différentes lois qui affectent les entreprises québécoises lorsque celles-ci désirent effectuer un projet impliquant des SIA. Plus précisément, nous désirons identifier les lois qui pourraient affecter directement l'algorithme qui sera créé afin d'obtenir un niveau d'explicabilité acceptable de celui-ci.

Puisque nous recherchons les lois qui pourraient affecter directement l'algorithme, cette recherche ne discutera pas des lois entourant la collecte, la conservation ou la suppression de données. Cette recherche se concentre essentiellement sur les lois applicables au Québec, c'est-à-dire, les lois provinciales et fédérales. Même si certaines entreprises québécoises pourraient être tenues de respecter des lois provenant d'autres provinces ou pays du fait de leurs activités internationales, cette recherche ne s'attardera que sur les lois applicables aux entreprises qui exploitent des SIA au Québec.

1.3.1 Législations québécoises portant sur les systèmes d'intelligence artificielle explicables

Au niveau provincial, une nouvelle loi encadrant l'utilisation des SIA a été adoptée, le 22 septembre 2021, nommée Loi modernisant des dispositions législatives en matière de

protection des renseignements personnels. Cette loi, qui entre graduellement en vigueur, au cours des années 2021 à 2024, a comme objectif d'offrir aux citoyens un meilleur contrôle de leurs renseignements personnels. Elle modernise le cadre législatif afin de l'adapter à la réalité technologique d'aujourd'hui (Gouvernement du Québec, 2022). Le traitement automatisé des renseignements personnels a été abordé par cette loi car il s'agit là d'une réalité technologique d'aujourd'hui qui est utilisée par plusieurs entreprises œuvrant au Québec. L'Article 12.1 de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels traite de sujets liés à la XIA.

« 12.1. Toute personne qui exploite une entreprise et qui utilise des renseignements personnels afin que soit rendue une décision fondée exclusivement sur un traitement automatisé de ceux-ci doit en informer la personne concernée au plus tard au moment où elle l'informe de cette décision.
Elle doit aussi, à la demande de la personne concernée, l'informer :
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.
Il doit être donné à la personne concernée l'occasion de présenter ses observations à un membre du personnel de l'entreprise en mesure de réviser la décision » (Gouvernement du Québec, 2021a).

Selon l'article 12.1 al. 3 de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels, les clients auront le droit de faire rectifier leurs renseignements personnels jugés incorrects ou obsolètes. Si ces renseignements avaient été utilisés dans la phase d'apprentissage du SIA, cela aurait certainement un impact sur toutes les prédictions émises par le système. Cependant, obliger les entreprises à changer les données des algorithmes pourrait nuire aux résultats livrés et par conséquent compromettre les performances de l'algorithme pour les autres utilisateurs. Par ailleurs, l'article 12.1 al. 2 de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels vient ajouter une tâche aux compagnies québécoises. Celle-ci pourrait, en fonction

de l'interprétation que l'on en fait, faire référence au droit à l'explication. Dans ce cas, les compagnies québécoises auront à donner à leurs clients une justification de la décision qui aura été prise de manière complètement automatisée (sans révision de la part d'un humain).

Il est à noter que l'utilisation d'un vocabulaire imprécis tel que « les principaux facteurs et paramètres » laisse une grande place à l'interprétation de la loi aux entreprises québécoises. Selon Jules Gaudin[2], l'utilisation de ces termes nébuleux permet aux entreprises de développer leurs propres techniques pour faire de l'IA explicable sans être restreintes dans les méthodes que celles-ci pourraient vouloir utiliser. Lors des débats de la Commission des institutions, le 27 mai 2021, Gaétan Barrette[3] a soulevé ce manque de clarté émis par cette loi. À la suite de cette intervention, Éric Caire[4] a clarifié ce qu'il entendait par cette formulation.

« Le paramètre, en langage informatique, c'est l'information que vous passez à l'algorithme pour qu'il puisse faire son traitement. Donc, la mécanique demeure interne, c'est l'information que vous voulez voir traiter par l'algorithme que vous lui donnez » (Assemblée nationale du Québec, 2019).

On peut alors suggérer que, même si cette clarification n'a pas été ajoutée dans la loi, c'est probablement cette avenue que les entreprises suivront afin de s'assurer d'être conformes à l'article 12.1 al. 2 de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels. Du point de vue technique, le niveau de difficulté de cette tâche est très faible ; cela est facilement réalisable pour la majorité des entreprises québécoises et ne nécessitera pas une grande phase d'adaptation. Il est important de noter, puisqu'il s'agirait

---

[2] Avocat en technologies émergentes et PI chez ROBIC, S.E.N.C.R.L. / LLP, maître en science politique et gouvernement, maître en droit des affaires, maître en droit des affaires comparés, maître en droit et en droit des affaires dans un contexte mondial.

[3] Député de La Pinière pour le Parti Libéral du Québec, membre du Bureau de l'Assemblée nationale, porte-parole de l'opposition officielle en matière d'accès à l'information, porte-parole de l'opposition officielle en matière d'éthique, porte-parole de l'opposition officielle en matière de justice.

[4] Député de La Peltrie pour la coalition avenir Québec, ministre de la Cybersécurité et du Numérique, Ministre responsable de l'Accès à l'information et de la Protection des renseignements personnels, leader parlementaire adjoint du gouvernement.

uniquement de fournir les données d'entrées à l'algorithme, aucune explication sur la logique de celui-ci ne serait requise, selon la réglementation québécoise en vigueur.

Cependant, il est tout de même important de mentionner qu'en ingénierie, les principaux facteurs et paramètres n'ont pas la même définition que celle fournie par Éric Caire. En ingénierie informatique, les définitions de ses termes abordent des éléments qui ont trait à la logique de l'algorithme. Au Québec, il y a donc une discordance entre ce que les experts en ingénierie pourraient comprendre de cette loi et ce qui est réellement attendue de la part des législateurs.

1.3.2 Législations canadiennes portant sur les systèmes d'intelligence artificielle explicables

Dans le secteur public, le gouvernement fédéral s'est engagé à respecter, ainsi que tous ses fournisseurs, les Principes directeurs en matière d'intelligence artificielle qui stipulent le Principe 3 comme suit :

« Pour assurer une utilisation efficace et éthique de l'intelligence artificielle, le gouvernement veillera à :
3. Fournir des explications claires sur le processus décisionnel en matière d'intelligence artificielle tout en offrant des occasions d'examiner les résultats et de remettre en question les décisions » (Gouvernement du Canada, 2018).

Dans le secteur privé, le Projet de Loi C-11, nommé Loi de 2020 sur la mise en œuvre de la Charte du numérique, propose des pistes qui permettraient de moderniser la Loi sur la protection des renseignements personnels et documents électroniques. L'une des conclusions énoncées par le gouvernement est qu'il convient de redéfinir le concept de transparence dans la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDÉ) pour imposer une obligation d'informer toute personne non seulement de l'utilisation des processus automatisés et des facteurs influant la décision, mais aussi de son impact et de la logique derrière celle-ci (Innovation, Sciences et Développement économique

Canada, 2019). Cependant, du fait du climat politique incertain des élections fédérales en 2021 et plusieurs questionnements sur la légitimité du pouvoir fédéral sur le Projet de Loi C-11 (possible empiètement des compétences provinciales), ce projet a été mis sur pause.

Le 15 juin 2022, le Projet de Loi C-27, aussi connu sous le nom de Loi de 2022 sur la mise en œuvre de la Charte du numérique, a succédé au Projet de Loi C-11. L'intitulé complet du nouveau projet de loi est Projet de Loi C-27 : Loi édictant la Loi sur la protection de la vie privée des consommateurs (LPVPC), la Loi sur le Tribunal de la protection des renseignements et des données et la Loi sur l'intelligence artificielle et les données (LIAD). Le Projet de Loi C-27 réintroduit deux lois précédemment contenues dans le Projet de Loi C-11 (2020) : Loi relative à la protection des données, ainsi que la Loi sur la vie privée des consommateurs. La principale nouveauté du Projet de Loi C-27 tient dans l'introduction d'une troisième loi, la LIAD. Si celle-ci est adoptée, elle deviendrait la première loi au Canada à réguler spécifiquement l'utilisation des SIA. Selon le Projet de Loi C-27, l'objectif de la LIAD est d'établir des exigences nationales communes pour la conception, le développement et le déploiement de SIA conformes aux normes nationales et internationales. Selon Maître Chassigneux[5], ce projet de loi s'inspire d'initiatives similaires, telles que le Règlement général sur la protection des données de l'Union européenne. En revanche, le Projet de Loi C-27 pourrait donner lieu, en cas de poursuite, à des amendes ayant été décrites parmi les plus sévères du G7. Il est donc primordial que les entreprises se préparent avant la promulgation de la loi et ainsi éviter de fortes amendes.

L'explicabilité et la transparence des SIA ont été présentées dans la LPVPC ainsi que dans la LIAD, toutes deux faisant partie du Projet de Loi C-27. La LPVPC, tout comme la loi québécoise traitant des renseignements personnels, abord de thèmes relatifs à l'explicabilité.

[5] A travaillé pendant plus de dix ans à la Commission d'accès à l'information du Québec, dont six ans à titre de juge administratif affecté à la section de surveillance.

« (1) Sur demande de l'individu, l'organisation lui indique si elle détient des renseignements personnels qui le concernent, quel est l'usage qu'elle en fait et si elle les a communiqués. Elle les met à sa disposition.

Nom des tiers ou catégories de tiers

(2) Si l'organisation a communiqué les renseignements, elle fournit à l'individu le nom des tiers ou les catégories de tiers auxquels ils ont été communiqués, et ce, même lorsqu'elle les a communiqués sans son consentement.

Système décisionnel automatisé

(3) Si l'organisation a utilisé un système décisionnel automatisé pour faire une prédiction, formuler une recommandation ou prendre une décision concernant l'individu et que la prédiction, la recommandation ou la décision pourrait avoir une incidence importante pour lui, elle lui en fournit, à sa demande, une explication.

Explication

(4) L'explication indique le type de renseignements personnels utilisés pour faire la prédiction, formuler la recommandation ou prendre la décision, la provenance de ces renseignements ainsi que les motifs ou les principaux facteurs ayant mené à la prédiction, à la recommandation ou à la décision » (Gouvernement du Canada, 2022c).

La définition des termes « incidence importante » n'a pas encore été spécifiée. Selon Alvavi *et al.* (2022), nous pouvons supposer que cette notion inclura des circonstances susceptibles d'entraîner un préjudice grave, puisque ces termes sont définis ainsi à l'article 58(7) de la LPVPC. Tout comme la Loi 64 (loi québécoise), on retrouve dans ce texte de loi des termes nébuleux, soit, l'explication des motifs ou les principaux facteurs ayant mené à la prédiction, à la recommandation ou à la décision. Par contre, il est important de souligner que ce projet de loi n'a pas encore été adopté et est susceptible de changements. Il est possible que les termes « motifs et principaux facteurs » soient définis dans une future version de ce projet loi, ce qui permettrait de clarifier les lois fédérale et provinciale.

Tout comme le provincial, le projet de loi fédéral qui aborde l'explicabilité des SIA utilise des termes nébuleux. Dans l'état actuel, il y a donc aussi une discordance entre ce que les experts

en ingénierie pourraient comprendre de cette loi et ce qui est réellement attendue de la part des législateurs.

1.4 Les systèmes d'intelligence artificielle explicables et éthiques

Si l'éthique et la robustesse des SIA sont dans une certaine mesure déjà reflétées dans la législation existante, leur pleine réalisation pourrait dépasser les obligations juridiques existantes. En effet, la législation peut ne pas toujours suivre les évolutions technologiques, ne pas correspondre aux normes éthiques ou s'avérer inadaptée à certaines situations. Il est donc nécessaire que les systèmes d'IA soient également conformes aux normes éthiques pour être considérés dignes de confiance. L'éthique en matière d'IA se définit comme suit :

« L'éthique en matière d'intelligence artificielle est un sous-domaine de l'éthique appliquée qui est axé sur les questions d'ordre éthique soulevées par la mise au point, le déploiement et l'utilisation de l'intelligence artificielle. Sa préoccupation centrale consiste à déterminer la manière dont l'intelligence artificielle peut soulever des préoccupations relatives au bien-être des individus ou y apporter des solutions, que ce soit du point de vue de la qualité de vie ou de l'autonomie humaine et de la liberté nécessaire pour une société démocratique » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

Dans les sections suivantes, nous allons d'abord nous intéresser à l'essor de l'éthique de l'IA et son influence croissante sur la technologie québécoise. À la section 1.4.1, nous abordons les principes fondamentaux de l'IA éthique établis par la Déclaration de Montréal pour un développement responsable de l'IA, particulièrement ceux qui traitent des thématiques de XIA. Ensuite, à la section 1.4.2, nous cherchons à identifier comment les ingénieurs peuvent prendre ces principes éthiques et les appliquer dans un cadre pratique. La section 1.4.3 propose des réflexions et des pistes de solutions pour faire avancer la littérature sur l'IA éthique.

1.4.1 L'essor de l'éthique en intelligence artificielle

En 2022, nous pouvons compter plus de 80 codes d'éthique portant sur l'IA (Munn, 2023). Ceux-ci comportant tous un contenu similaire, il est juste de dire que tous s'entendent pour dire que la transparence doit être respectée lors du développement de SIA. L'une des initiatives québécoises, reconnue internationalement, est La déclaration de Montréal pour un développement responsable de l'intelligence artificielle (DDIRIA).

« Il s'agit d'une œuvre collective qui a pour objectif de mettre le développement de l'IA au service du bien-être de tous et chacun, et d'orienter le changement social en élaborant des recommandations ayant une forte légitimité démocratique » (Université de Montréal, 2018a).

Cette déclaration met en avant la transparence et l'explicabilité comme une nécessité lors de l'implémentation d'IA responsable. Elle a été élaborée grâce à un processus participatif qui a impliqué la signature de plus de 10 ordres professionnels québécois. Au niveau des institutions académiques, ce sont aussi plus d'une dizaine d'universités, cégeps et écoles secondaires qui ont soutenu cette initiative. De plus, une quinzaine d'organismes, représentant l'autorité publique québécoise ou la conseillant, ont signé la DDIRIA. La liste de signataires comprend aussi plusieurs organismes fédéraux, centres de recherches québécois et canadiens reconnus mondialement pour leurs recherches en IA, tel que le Conseil national de recherches du Canada (CNRC). Des institutions créditées pour leurs connaissances dans le domaine de l'éthique, tel que l'association des praticiens en éthique du Canada (APEC), font également partie des signataires.

La DDIRIA présente 10 principes éthiques pour un développement responsable de l'IA : le bien-être, le respect de l'autonomie, la protection de l'intimité et de la vie privée, la solidarité, la participation démocratique, l'équité, l'inclusion de la diversité, la prudence, la responsabilité et finalement, le développement soutenable. Chacun de ces principes est accompagné de directives qui doivent être respectées afin que le développement de l'IA soit conforme au principe. Tel que mentionné précédemment, l'utilisation de systèmes intelligents explicables

est une partie du casse-tête qui permet de créer une IA de confiance. Il est donc normal de s'attendre à ce que l'utilisation d'une IA explicable permette de répondre à certaines directives de cette charte. Voici quelques-unes de ces directives.

« Les SIA doivent satisfaire les critères d'intelligibilité, de justifiabilité et d'accessibilité, et doivent pouvoir être soumis à un examen, un débat et un contrôle démocratiques.
1) Le fonctionnement des SIA qui prennent des décisions affectant la vie, la qualité de la vie ou la réputation des personnes doit être intelligible pour leurs concepteurs.
2) Les décisions des SIA affectant la vie, la qualité de la vie ou la réputation des personnes, devraient toujours être justifiables dans un langage compréhensible aux personnes qui les utilisent ou qui subissent les conséquences de leur utilisation. La justification consiste à exposer les facteurs et les paramètres les plus importants de la décision et doit être semblable aux justifications qu'on exigerait d'un être humain prenant le même type de décision » (Université de Montréal, 2018b).

Il ne serait donc pas faux, à la suite de ces observations (nombre de signataires et directives de la charte), de dire que l'industrie canadienne souhaite créer des SIA qui sont transparents et explicables.

Finalement, ce qui pose souvent problème dans l'industrie, ce n'est pas la reconnaissance des principes éthiques, mais la mise en place de mesures concrètes pour les appliquer. D'ailleurs, dans la littérature portant sur l'IA de confiance, nous pouvons fréquemment trouver l'expression suivante : « Nous n'en sommes plus à définir le quoi, mais plutôt le comment ». Dans cette affirmation, le « quoi » signifie les principes et les faits moraux qui peuvent orienter de manière éthique l'innovation numérique, en particulier en matière d'IA, pour le bénéfice de l'humanité (Floridi, 2019). Le « comment » définit les techniques d'intégration de l'éthique de manière concrète et efficace pour obtenir un impact bénéfique (Floridi, 2019).

1.4.2 Passer du « quoi » au « comment »

De plus en plus de chercheurs en éthique de l'IA, tels que Munn (2023), s'interrogent sur l'efficacité de l'établissement de codes éthiques sans disposer de moyens concrets pour garantir leur respect. Munn argumente que les principes éthiques sont considérés comme un échec puisque ceux-ci manquent de mesures concrètes. Du point de vue de la pratique des ingénieurs, s'en tenir à une description de grands principes devant régir les technologies n'est pas éclairant. Il faut relier ces principes à des mesures concrètes.

Il y a différentes manières de concevoir la relation entre, d'une part, les valeurs que l'on souhaite mettre de l'avant dans les systèmes d'IA, et d'autre part les mesures concrètes pour y parvenir. Selon la définition d'une approche universaliste, développer des standards et mesures applicables à tous les contextes, tel que discuté dans Reddy et al. (2023), Weber et al. (2023) ou encore Le et al. (2023), en mettant l'accent, par exemple, sur la création d'un lexique commun, de méthodes uniformes, et de critères d'évaluation fixes pour les décisions des modèles d'IA permettrait de fournir aux ingénieurs des mesures concrètes en XIA. Au contraire, les approches contextuelles adaptent l'explicabilité aux spécificités des différents domaines d'application, reconnaissant que les besoins et les définitions de l'explicabilité varient selon le contexte ciblé. Des recherches telles que Nyrup et Robinson (2022) ainsi que Nauta et al. (2023) explorent l'approche contextuelle de la XIA. Ainsi, les mesures concrètes pour répondre aux enjeux éthiques dépendraient des faits de la situation (par exemple : les types d'utilisateurs de la plateforme et l'environnement dans lequel la solution est déployée).

Chercher des mesures concrètes universelles est tentant. Après tout, ces solutions ont le mérite de « durer dans le temps », et de s'appliquer à une foule de situations. Elles permettent d'établir une base commune de mesures nécessaires ou suffisantes pouvant être importées dans toutes les organisations ou tous les systèmes. C'est pourquoi de nombreux programmes de recherche ont pour objectif d'identifier de telles mesures. Or, est-il réaliste d'espérer identifier de telles mesures concrètes universelles pour la XIA? La normativité offre un cadre pour passer du « quoi » au « comment » qui permettrait d'explorer cette question.

La normativité guide l'industrie non seulement dans l'élaboration de normes éthiques, mais aussi dans la mise en place de mesures concrètes pour assurer que ces normes soient intégrées efficacement dans les processus technologiques. Dans le domaine de l'IA, des approches normatives ont été utilisées afin d'évaluer l'efficacité de mesures concrètes pour contrer les biais. Nous pouvons citer en exemple Larsson (2019), Ferrer et al. (2021) ainsi que Kostick-Quenet et al. (2023). En revanche, les recherches qui incorporent des approches normatives pour l'évaluation de mesures concrètes en XAI sont limitées. L'importance de la normativité en XIA réside dans son rôle de guide pour le développement de technologies qui respectent, non seulement les exigences techniques, mais aussi les valeurs éthiques, renforçant ainsi la confiance et l'acceptabilité sociale de l'IA.

Par conséquent, l'application d'évaluations normatives de mesures concrètes dans le domaine de la XIA permettrait de confirmer ou d'infirmer l'utilité des approches contextuelles, en plus de fournir un cadre de définition de mesures concrètes adaptés aux connaissances des experts dans le domaine de l'IA. Malheureusement, le manque de travaux dans la littérature qui traite de ce sujet ne nous permet pas d'identifier une approche pour passer du « quoi » au « comment » dans le domaine spécifique de la XIA. Les ingénieurs sont donc tenus responsable de systèmes d'IA qualifiés d'explicables et transparents, sans être en mesure d'évaluer si ceux-ci répondent réellement à leurs valeurs éthiques.

1.4.3 Perspectives et recommandations

Pour conclure cette section dédiée aux SIA explicables et éthiques, il est important de souligner que la majorité des acteurs dans le domaine s'accordent sur un point crucial : pour être éthique, une IA doit être explicable. Cette reconnaissance est attestée par l'engagement des entreprises, des industries et des ordres professionnels québécois qui ont signé la DDIRIA. Cette adhésion reflète une volonté de développer des systèmes intelligents artificiels (SIA) qui sont non seulement éthiques mais aussi transparents et explicables.

Cependant, bien que l'intention soit claire, l'industrie se heurte à des difficultés significatives dans l'implémentation pratique de ces principes éthiques dans le développement des SIA. Un des défis majeurs réside dans l'identification et l'application de mesures concrètes appropriées. Actuellement, il existe une certaine incertitude quant à la sélection des techniques d'explicabilité à utiliser, ainsi que sur le moment et le contexte spécifique de leur application dans divers projets.

Face à cette problématique, il devient évident que l'industrie québécoise, ainsi que le domaine de l'IA en général, bénéficieraient grandement d'un guide, d'une méthodologie ou d'une procédure structurée permettant de naviguer au travers les subtilités des contextes et les différentes applications de l'IA. Un tel outil offrirait un cadre clair et des directives précises pour déterminer quelle technique d'explicabilité devrait être mise en œuvre et dans quel contexte spécifique. Ceci afin de répondre efficacement aux exigences éthiques et pratiques. L'élaboration de cette méthodologie contribuerait, non seulement à la mise en pratique des principes éthiques, mais aussi à l'avancement de l'IA en tant que technologie socialement responsable et digne de confiance.

1.5 Les systèmes d'intelligence artificielle explicables et robustes

Les SIA continuent de gagner en sophistication, mais ils peuvent être confrontés à des défis lorsqu'ils sont exposés à des environnements inconnus. Dans ces contextes, la robustesse de l'IA devient essentielle pour garantir un fonctionnement fiable et sécurisé. Cette caractéristique cruciale permet aux SIA de maintenir leur performance, quels que soient les scénarios et les situations auxquels ils sont confrontés. La robustesse de l'IA étant une caractéristique de l'IA de confiance, il convient de définir comment la XIA contribue aux prouesses technologiques de celle-ci. Selon la norme ISO/IEC TR 24029-1, un SIA robuste possède l'habilité de maintenir le même niveau de performance, peu importe les circonstances (International Organization for Standardization, 2021).

Dans la section suivante, nous nous intéressons à la manière dont la XIA s'intègre dans la recherche sur la robustesse de l'IA. Pour ce faire, deux perspectives, ayant des objectifs différents, sont abordées : utiliser la XIA pour démontrer la robustesse du système dans son ensemble (section 1.5.1) et, assurer que les explications sont, en elles-mêmes, robustes (section 1.5.2).

D'une part, les techniques de la XIA permettent de vérifier, démontrer et améliorer la robustesse des SIA. Les équipes qui créent les SIA ont une meilleure compréhension de ceux-ci lorsqu'une explication de la logique utilisée pour faire des prédictions leur est fournie. Cette meilleure compréhension mène à une identification plus rapide des erreurs de modélisations qui pourraient affecter la robustesse de l'IA. Ainsi, grâce aux explications fournies par les techniques de la XIA, il est possible d'évaluer et d'augmenter la robustesse d'un modèle et du système dans sa globalité. La section 1.5.1 explore différents protocoles d'expérimentation où la XIA a été utilisée afin de démontrer ou de renforcer la robustesse d'un SIA. Nous comprendrons que la XIA est utilisée avantageusement à cet effet. Néanmoins, le manque de protocoles est flagrant et est démontré par les démarches personnalisées, au cas par cas.

D'autre part, il convient également d'évaluer la robustesse des techniques de XIA elles-mêmes. Dans la section 1.5.2, nous démontrons qu'il n'est pas garanti que la logique décrite dans l'explication soit celle réellement utilisée par l'algorithme. Cet enjeu relève de la robustesse de l'explication. La section 1.5.2 explore aussi les différentes métriques qui permettent d'évaluer la robustesse des explications fournies par les techniques de XIA. Cette analyse met en évidence la nature encore expérimentale des métriques de performances de la XIA. Il ressort que l'identification des besoins en XIA d'un système permettrait de mieux définir et de prioriser les métriques d'évaluation de la XIA.

1.5.1 L'explicabilité d'un système d'intelligence artificielle pour augmenter la robustesse

L'explicabilité d'un SIA joue un rôle crucial dans l'amélioration de sa robustesse. Les algorithmes d'IA transparents sont souvent plus faciles à auditer et à comprendre, ce qui peut

conduire à une meilleure détection et correction des erreurs de modélisation. Cela renforce à son tour la robustesse du système. Ce facteur peut être démontré grâce à une multitude d'études et d'articles qui en font mention. Cette section examine les protocoles suivis par ces études afin de définir les forces et faiblesses des méthodologies d'évaluation de la robustesse des SIA grâce à la XIA.

Par exemple, Mahima *et al.* (2021) examinent le processus de prise de décision de leur système d'IA avec des techniques de XIA. Cette analyse leur a permis de comparer des modèles réalisant une même tâche et de choisir le plus performant en fonction des raisons sous-jacentes aux décisions produites. En évaluant si ces raisons étaient cohérentes, pertinentes et stables d'un cas à l'autre, ils ont pu juger de la robustesse du système, c'est-à-dire sa capacité à fournir des résultats fiables même dans des contextes variés. Ainsi, l'explicabilité a été utilisée comme un outil d'audit interne pour détecter les modèles dont le comportement était moins fiable ou plus sensible aux variations d'entrée.

Un autre cas est l'étude de Raz *et al.* (2022), qui utilise la XIA pour identifier les caractéristiques d'entrée ayant le plus d'influence sur les prédictions du modèle, dans le but de mieux comprendre les raisons sous-jacentes à ses décisions. À partir de cette liste de caractéristiques influentes, les auteurs mènent une évaluation partielle de la robustesse du système en comparant le comportement du modèle sur des exemples pour lesquels il a été entraîné (exemples connus), à celui observé sur des exemples qu'il n'a jamais vus auparavant (exemples non connus). Ils analysent notamment si le système s'appuie sur les mêmes types de caractéristiques pour prendre ses décisions, ce qui permet de juger de sa capacité à généraliser de manière cohérente, un indicateur important de robustesse.

En 2022, Pandianchery *et al.* ont également eu recours à des techniques d'explicabilité pour évaluer la robustesse de leur SIA. Leur modèle avait pour objectif de prédire le taux de contamination à la Covid-19 dans les différentes provinces de l'Inde. Pour tester sa capacité de généralisation, le système a été entraîné uniquement à partir des données d'une seule province, puis évalué sur des données provenant des autres provinces. À l'aide de méthodes

d'explicabilité, les chercheurs ont pu analyser les variables sur lesquelles le modèle fondait ses prédictions, et vérifier si les règles ou motifs appris étaient cohérents et applicables d'une province à l'autre. Cette démarche leur a permis de conclure que le modèle ne s'était pas limité à des corrélations spécifiques à une région, mais avait appris des relations plus générales et transférables (un indicateur de robustesse).

Alors que l'utilisation de la XIA dans les études mentionnées (par exemple celles de Mahima *et al*., Raz *et al.*, et Pandianchery *et al.*) vise à améliorer la robustesse des SIA, les techniques utilisées pour y parvenir varient considérablement selon les cas d'usage. Cette variation s'explique par plusieurs facteurs, notamment le type de données utilisées (données textuelles, images ou données tabulaires), ainsi que la nature des tâches assignées aux modèles (classification, régression, apprentissage par renforcement, etc.). Certaines études ont eu recours à des outils open source pour l'explicabilité, tandis que d'autres ont développé leurs propres méthodes adaptées à leur contexte. Ces travaux n'ont pas défini de protocoles standardisés ni de bancs d'essai d'explicabilité en amont ; ils ont plutôt appliqué les techniques de XIA directement pour évaluer leurs résultats. De plus, les approches explicatives varient dans leur logique sous-jacente, bien qu'elles visent toutes un objectif commun : démontrer la robustesse du système étudié.

Contrairement aux caractéristiques telles que la légalité ou l'éthique, qui concernent principalement la relation entre le système et les utilisateurs finaux, l'explicabilité appliquée à la robustesse cible plutôt les professionnels impliqués dans le développement ou l'évaluation des SIA. Puisque ces experts possèdent généralement une bonne compréhension technique des modèles, des explications centrées sur l'impact des caractéristiques d'entrée ou sur les règles décisionnelles internes, la XIA se révèlent particulièrement utiles pour soutenir leurs analyses.

À la lumière des études analysées, nous supposons que la robustesse d'un système d'IA peut être vérifiée, démontrée et même renforcée au moyen de techniques d'explicabilité (XIA), telles que l'analyse de l'importance des caractéristiques d'entrée (ex. SHAP, LIME), la visualisation des activations internes du modèle, ou encore l'interprétation de règles logiques

extraites des prédictions. En revanche, il n'existe actuellement aucun protocole ou cadre méthodologique standardisé pour guider les développeurs souhaitant utiliser ces techniques dans une optique d'évaluation de la robustesse. À ce jour, les approches sont conçues au cas par cas, chaque équipe adaptant des techniques de XIA spécifiques à son contexte, à ses données ou à ses objectifs. Cela rend leur réutilisation difficile, voire impossible, dans d'autres projets.

Cette absence de standardisation a également été signalée dans la littérature, notamment par Saeed et Omlin (2023). Ces auteurs constatent que les travaux en XIA sont souvent menés de manière indépendante les uns des autres, sans coordination ni mutualisation des résultats. Ils soulignent ainsi la nécessité de consolider les méthodes développées jusqu'à présent, afin de construire des protocoles génériques et réutilisables, qui pourraient être appliqués à différents types de systèmes et de cas d'usage.

1.5.2 La robustesse d'une explication d'un système d'intelligence artificielle

Un aspect essentiel du lien entre robustesse et explicabilité réside dans le fait que la robustesse constitue une caractéristique nécessaire à l'explicabilité. En fait, l'explicabilité d'un algorithme doit elle-même respecter les exigences de robustesse.

Un parallèle avec la définition d'une IA robuste nous permet de définir la notion d'explication robuste. Une explication robuste possède l'habilité de maintenir le même niveau de performance, peu importe les circonstances. Cette définition implique deux éléments importants à considérer : la capacité de caractériser les performances d'une explication, mais aussi la capacité de vérifier les performances au travers d'une multitude de circonstances.

Les expériences de Neely *et al.* (2021) montrent que l'accord entre différentes méthodes de XIA est souvent faible, en particulier lorsqu'elles sont appliquées à des modèles complexes de type boîte noire. De leur côté, Jyoti *et al.* (2022) recensent plusieurs travaux récents ayant mis en évidence que les explications générées (par exemple l'importance attribuée aux

caractéristiques d'entrée, les cartes de chaleur ou les règles décisionnelles extraites) peuvent varier considérablement lorsqu'on introduit de légères perturbations dans les données d'entrée d'un SIA. Ces observations mettent en lumière un déficit de robustesse des approches de la XIA, dans la mesure où les résultats explicatifs ne sont pas fiables : ils manquent de stabilité et peuvent induire en erreur les parties prenantes qui s'y fient pour comprendre ou valider le comportement du système d'IA.

Les recherches de Jyoti *et al.* (2022) ont établi que malgré le nombre croissant de techniques d'explicabilité proposées dans la littérature, peu de recherches s'aventurent dans la définition d'un protocole d'évaluation de ces techniques, en raison de la subjectivité des interprétations et du manque d'explications de référence. Afin d'aborder ces problématiques sous une nouvelle perspective, une grande quantité d'études ont d'abord voulu définir les propriétés d'une bonne explication. Nauta *et al.* (2023) ont effectué une revue de littérature, composée de plus de 600 articles, portant sur les métriques d'évaluation de la XIA. Ils ont suggéré que l'explicabilité est un élément non-binaire mesurant le degré de satisfaction de certaines propriétés. En regroupant les différentes terminologies présentes dans la littérature et en minimisant leur chevauchement, Nauta *et al.* (2023) déterminent les propriétés d'une bonne explication dans le domaine d'expertise de la XIA, telles que présentées dans le tableau 1.1.

En plus de regrouper les différentes propriétés d'une bonne explication, Nauta *et al.* (2023) ont aussi répertorié les différentes techniques d'explicabilité. Les plus communes, ordonnancées par popularité, sont les suivantes : utilisation d'une carte thermique, présentation de l'importance des caractéristiques, localisation, explications textuelles, présentation de prototypes, désentrelacement, graphiques, règles de décisions, création d'une synthèse de représentations, visualisation de représentations, création d'un modèle de boîte blanche, présentation d'un graphique des caractéristiques et utilisation d'un arbre de décision. Dans un contexte idéal, l'évaluation de chaque propriété d'une explication de qualité devrait être possible pour chacune de ces techniques d'explicabilité de l'IA. Il convient donc de s'intéresser aux métriques permettant de quantifier ces propriétés.

Tableau 1.1 Propriétés d'une bonne explication, regroupées par dimension,
Tiré de Nauta *et al.* (2023)

Dimension	**Propriété**	**Description**
Contenu	Exactitude	Décrit dans quelle mesure l'explication est fidèle à la boîte noire.
	Complétion	Décrit dans quelle mesure le comportement de la boîte noire est représenté dans l'explication.
	Consistance	Décrit dans quelle mesure la méthode d'explication est déterminante et invariante dans l'implémentation.
	Continuité	Décrit dans quelle mesure la fonction d'explication est continue et généralisable.
	Contraste	Décrit dans quelle mesure l'explication est discriminante par rapport à d'autres événements ou cibles.
	Complexité de covariation	Décrit la complexité des (interactions des) caractéristiques dans l'explication.
Présentation	Compacité	Décrit la taille de l'explication.
	Composition	Décrit le format de présentation et l'organisation de l'explication.
	Confiance	Décrit la présence et l'exactitude des informations de probabilité dans l'explication.
Utilisateur	Contexte	Décrit la pertinence de l'explication pour l'utilisateur et ses besoins.
	Cohérence	Décrit dans quelle mesure l'explication est conforme aux connaissances et croyances antérieures.
	Contrôlabilité	Décrit dans quelle mesure une explication est interactive ou contrôlable pour un utilisateur.

Il n'y a pas de méthodes d'évaluation largement acceptées pour la XIA, contrairement aux métriques d'évaluation standards existantes pour estimer la performance de l'IA. Par conséquent, il est courant d'observer des méthodes d'évaluation de la XIA qui ne sont pas basées sur des métriques quantifiables, mais plutôt sur des expériences individuelles où l'intuition du chercheur confirme la validité de l'explication. Nauta *et al.* (2023) mentionnent que 33% des études en XIA évaluent les performances uniquement basées sur des évidences anecdotiques individuelles, 58% des études appliquent une forme d'évaluation quantitative et 22% des études font une évaluation en demandant le retour de personnes dans l'étude des

utilisateurs. Sur ces 22%, 23% des études estiment la performance de la XIA en faisant appel uniquement à des experts IA.

Les pratiques d'évaluation de la XIA ont varié au courant des dernières années. On observe que les publications évaluant des techniques de XIA de manière quantitative ont légèrement augmenté, alors que les évaluations basées sur des évidences anecdotiques ont diminué. En revanche, le nombre de publications impliquant des études utilisateur est resté stable, aux alentours de 20%. Ce nombre est très bas, considérant que la majorité des recherches en XIA s'entendent pour dire que les besoins utilisateurs devraient guider le domaine de la XIA (Haque et al., 2023). La figure 1.3 présente cette évolution des mentalités dans le domaine de l'évaluation de la XIA.

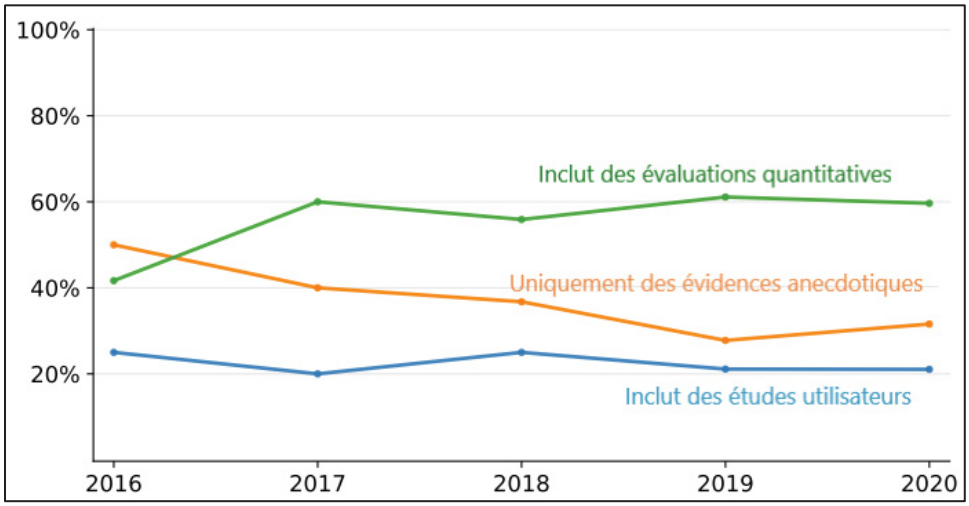


Figure 1.3 Évolution des techniques d'évaluation de la XIA entre les années 2016 et 2020, Adaptée de Nauta *et al.* (2023)

Il a été reconnu qu'une évaluation quantitative de la XIA permettrait des comparaisons intéressantes entre celles-ci (Batic et al., 2023). En suivant cette optique, nous avons répertorié, dans le tableau 1.2, les propriétés qui peuvent être évaluées grâce à des métriques d'évaluation quantitative pour les techniques les plus utilisées en XIA. De ce tableau, on peut conclure qu'il n'existe pas de métriques permettant d'évaluer chacune des propriétés de chacune des

techniques de XIA. Il est important de spécifier que toutes les métriques ont été répertoriées alors même que l'étude de Jyoti *et al.* (2022), montre qu'elles ne sont pas idéales.

Tableau 1.2 Propriétés pouvant être évaluées quantitativement sans impliquer un utilisateur, pour les méthodes existantes de XIA

	Exactitude	Complétion	Consistance	Continuité	Contraste	Complexité de covariation	Compacité	Composition	Confiance	Contexte	Cohérence	Contrôlabilité
Carte thermique	X	X		X	X	X	X				X	
Importance des caractéristiques	X	X	X	X	X	X	X		X	X	X	
Localisation	X	X		X	X	X	X			X	X	
Texte		X		X	X		X	X			X	X
Prototypes	X	X		X		X	X		X		X	
Désentrelacement					X	X					X	
Graphique	X			X			X					
Règles de décision	X	X		X		X	X			X		
Synthèse de représentation				X	X	X	X	X		X		
Visualisation de représentation												
Modèle en boîte blanche	X	X		X			X					
Graphique des caractéristiques												
Arbre de décision		X					X					

Plusieurs problématiques issues de ces métriques sont identifiées. Un des problèmes mis en relief est la contradiction qui peut exister face aux différents aspects des explications. C'est pourquoi, l'identification et la priorisation des besoins en XIA sont primordiales afin de bien déterminer les métriques d'évaluation des techniques de XIA permettant la résolution d'une

problématique spécifique. En outre, les métriques d'évaluation de l'IA explicable sont encore largement expérimentales et n'ont pas encore été utilisées dans de nombreuses études. Il est donc important de prendre en compte les spécificités de chaque contexte d'application avant de choisir les métriques d'évaluation les plus appropriées.

Toutes ces problématiques ont poussé certains chercheurs à s'interroger sur l'évaluation des techniques de la XIA en ayant un regard axé sur la philosophie des sciences, la psychologie, le facteur humain et même d'un point de vue éducatif. Saeed et Omlin (2023) ont établi que des approches provenant de l'expertise psychologique permettent d'extraire la structure et les caractéristiques d'une explication afin de saisir comment celle-ci influence une personne. Plusieurs de ces recherches sur l'interaction homme-machine ont permis de comprendre que des études multidisciplinaires sont essentielles dans les avancements significatifs du domaine de la XIA (Saeed et Omlin, 2023). Cependant, peu d'expériences utilisateurs ont été menées dans ce domaine alors que plusieurs études, telles que celles de Jyoti *et al.* (2022), Suresh *et al.* (2021) ainsi que Arya *et al.* (2019), appuient explicitement sur l'importance de la prise en considération des aspects multidisciplinaires. À la suite d'une revue de la littérature, Shane *et al.* (2019) ont mis en évidence que les travaux existant sur l'évaluation de la XIA n'ont pas pris en considération les connaissances, les objectifs et habiletés des utilisateurs. Dans l'intention de diminuer l'écart entre la recherche et la réalité, il est nécessaire d'énoncer clairement les objectifs et les buts afin de guider la création de tests utilisateurs pertinents et représentatifs des facteurs essentiels à évaluer (Shane *et al.*, 2019). À l'heure actuelle, il y a peu de protocoles pour identifier les besoins des utilisateurs en XIA, ce qui complexifie la priorisation des métriques d'évaluation.

Présentement, il est clair que nous cherchons encore les meilleures méthodes pour évaluer les performances des techniques de XIA. Or, définir une XIA robuste suppose d'abord la capacité à caractériser de manière fiable la qualité des explications produites, puis à vérifier leur stabilité dans divers contextes d'utilisation. Tant que cette première étape (caractérisation de la qualité) reste mal définie, il demeure impossible de tester ces explications de façon cohérente à travers différents scénarios, et donc d'en garantir la robustesse. Nous constatons que cet axe de recherche est encore en évolution et que plusieurs opportunités d'étude et de développement

s'offrent aux chercheurs qui s'y intéressent comme la création d'un protocole d'identification des besoins en XIA, ou encore la création de métriques d'évaluation s'assurant que les techniques de la XIA répondent aux besoins identifiés ou bien l'optimisation de l'IA en fonction des métriques de XIA, avec pour objectif d'augmenter la robustesse de la XIA et de l'IA.

1.6 Cycle de vie d'un système d'intelligence artificielle explicable

La recherche dans le domaine de la XIA est souvent concentrée sur les techniques algorithmiques et n'aborde pas suffisamment la nature pratique de l'explicabilité dans des contextes réels. Cette focalisation étroite peut conduire à des conceptions techniquement solides mais échouant dans la pratique parce qu'elles ne tiennent pas compte des contextes pratiques dans lesquels les explications sont nécessaires. En examinant les explications nécessaires à des stades spécifiques du cycle de vie de l'IA, il sera possible de mettre en évidence les besoins en explicabilité qui diffèrent au travers des phases de l'IA.

Actuellement, au sein de la communauté de l'apprentissage automatique (ML), il existe deux visions différentes pour améliorer la performance des SIA : l'IA centrée sur le modèle (Model-Centric AI) et l'IA centrée sur les données (Data-Centric AI) (Hamid, 2022). Ces approches affectent la forme du cycle de vie d'un SIA.

La section suivante explore ces deux visions afin de déterminer laquelle favorise davantage la transparence et l'explicabilité. Compte tenu de la prédominance des petites et moyennes entreprises (PME) dans l'écosystème québécois de l'IA, nous examinons également comment les cycles de vie du SIA peuvent être adaptés à ces entreprises. Ces analyses visent à démontrer que l'approche Data-Centric AI favorise la transparence des SIA et pourrait constituer une stratégie efficace pour la gestion du cycle de vie de l'IA au sein des PME.

1.6.1 L'approche Model-Centric AI et ses limitations pour la XIA

L'approche Model-Centric AI (MCAI) représente une méthodologie dominante dans le développement de l'intelligence artificielle, mettant l'accent sur l'optimisation des modèles en termes de précision, de performance et d'efficacité computationnelle. Cette approche a façonné la manière dont l'IA a été abordée dans le domaine de la recherche et de l'industrie pendant des décennies (Hamid, 2022).

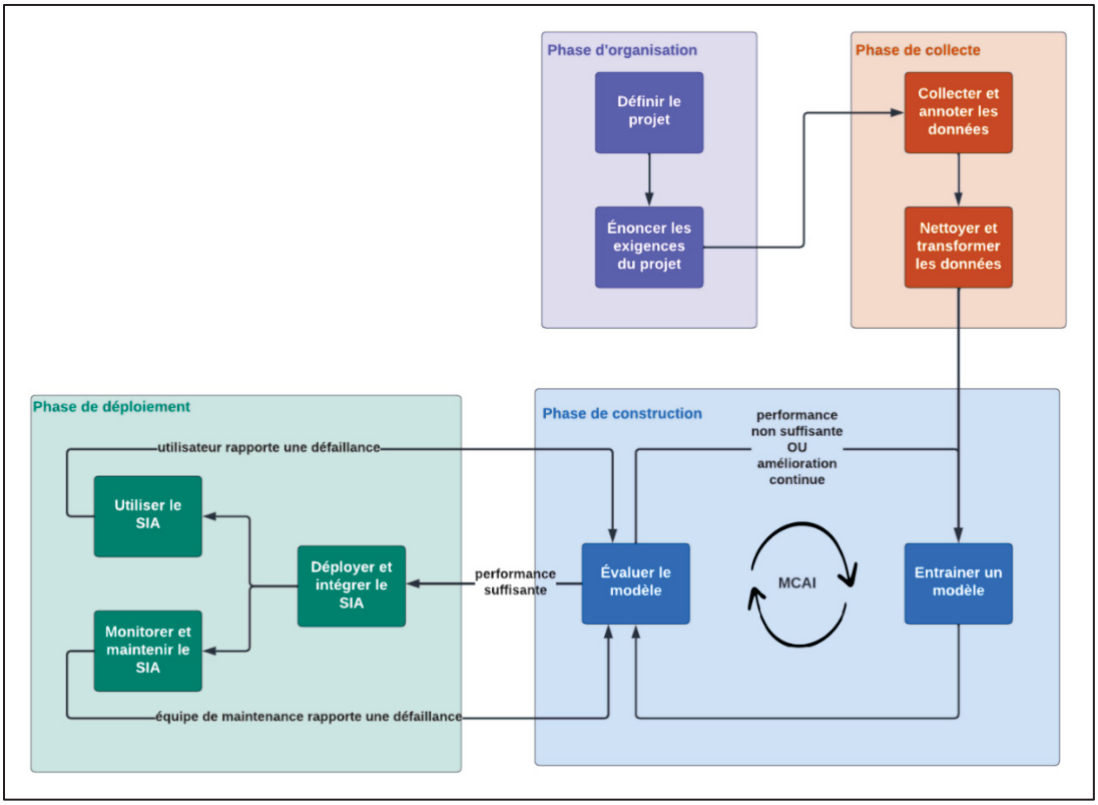


Figure 1.4 Cycle de vie d'un système d'intelligence artificielle incorporant l'approche Model-Centric AI

Historiquement et jusqu'à récemment, l'approche MCAI a dominé tant dans la recherche que dans l'industrie. Ng (2021) souligne que plus de 90% des projets de recherche en IA publiés ont utilisé une approche centrée sur le modèle. Dans cette approche, les données sont créées presque une seule fois et restent inchangées tout au long du cycle de vie du développement du système d'IA.

Malgré son succès au cours des trois dernières décennies, l'approche MCAI présente une limitation importante (Hamid, 2022). L'un des principaux inconvénients de cette méthode est la complexité croissante des modèles d'IA. Cette complexité peut obscurcir le fonctionnement interne des systèmes, rendant difficile pour les utilisateurs et les parties prenantes de comprendre comment les décisions sont prises par l'IA. Marcus et Davis (2019) souligne que cet accent sur les performances techniques, au détriment de la transparence et de la compréhensibilité, peut sérieusement entraver l'acceptation et la confiance dans les systèmes d'IA. Cette approche compromet l'explicabilité des systèmes d'IA.

Ainsi, bien que les modèles puissent atteindre des niveaux de performance élevés, leur manque d'explicabilité pose un problème majeur, en particulier dans des domaines où la compréhension des décisions prises par l'IA est cruciale.

1.6.2 Intégration de l'approche Data-Centric AI dans le cycle de vie des systèmes d'intelligence artificielle

Dans les dernières années, le Data-Centric AI (DCAI) a émergé comme un paradigme révolutionnaire, soulignant l'importance capitale de la qualité et de l'ingénierie des données dans la création de systèmes d'IA performants (Majeed et Hwang, 2023). À la différence de l'approche MCAI, qui se concentre principalement sur le perfectionnement du modèle d'IA lui-même, le DCAI met l'emphase sur l'amélioration des données, reconnaissant ainsi que la qualité des données est un facteur déterminant pour la performance d'un système d'IA. Ce paradigme favorise une collaboration étroite et itérative entre les experts du domaine qui apportent et interprètent les données, et les ingénieurs en machine learning, responsables de la construction et de l'affinement des modèles d'IA. Cette synergie entre les compétences et les connaissances assure un processus de développement plus intégré et efficace. En outre, le gouvernement canadien recommande l'utilisation du DCAI, notamment pour atténuer les risques associés à la gouvernance des données et pour promouvoir une utilisation responsable de l'IA (Office of the Superintendent of Financial Institutions Canada, 2023). L'avis du

gouvernement canadien souligne ainsi l'importance stratégique dans le contexte actuel de la technologie et de la réglementation.

L'approche Data-Centric AI est de plus en plus utilisée dans les grosses entreprises. On peut notamment citer Amazon qui applique l'approche DCAI pour personnaliser ses recommandations de produits et optimiser la gestion de son inventaire en fonction de la demande des clients (Charoliya, 2023). On peut également mentionner Waymo qui utilise l'approche DCAI pour former ses véhicules autonomes en utilisant des millions de miles de données de conduite (Charoliya, 2023).

Dans la dernière décennie, l'utilisation croissante de l'IA dans le milieu des entreprises a suscité une vague de recherches pour définir le cycle de vie optimal des SIA. Des études de cas menées par des entreprises de premier plan comme Microsoft (Amershi et al., 2019), Google (Ferlitsch, 2023) et IBM (Ishizaki, 2023) ont abouti à l'élaboration de modèles distincts de cycle de vie. Malgré des variations terminologiques et des niveaux de détail divergents, ces modèles partagent des similitudes fondamentales et ne se contredisent pas. Ils reflètent l'évolution des pratiques et l'importance croissante de l'IA dans le monde des affaires.

En intégrant l'approche DCAI aux principales phases d'un projet en IA, telles que définies par de grandes entreprises, nous aboutissons au cycle de vie présenté à la figure 1.5.

Bien que l'approche DCAI propose une amélioration significative dans la gouvernance des données et une définition du cycle de vie de l'IA, elle présente aussi des défis non négligeables. Des enquêtes révèlent que 96% des entreprises rencontrent des défis liés aux données, y compris à la qualité des données et à leur étiquetage dans les projets d'IA. 40% d'entre elles manquent de confiance dans leur capacité à assurer la qualité des données (Liang et al., 2022). Comme le DCAI est un concept relativement nouveau, de nombreux aspects critiques restent ambigus, notamment les définitions, les tâches associées, les algorithmes, les défis et les points de référence, ce qui peut entraver la progression vers des pratiques standardisées, particulièrement au sein de petites et moyennes entreprises (ZHA, 2023).

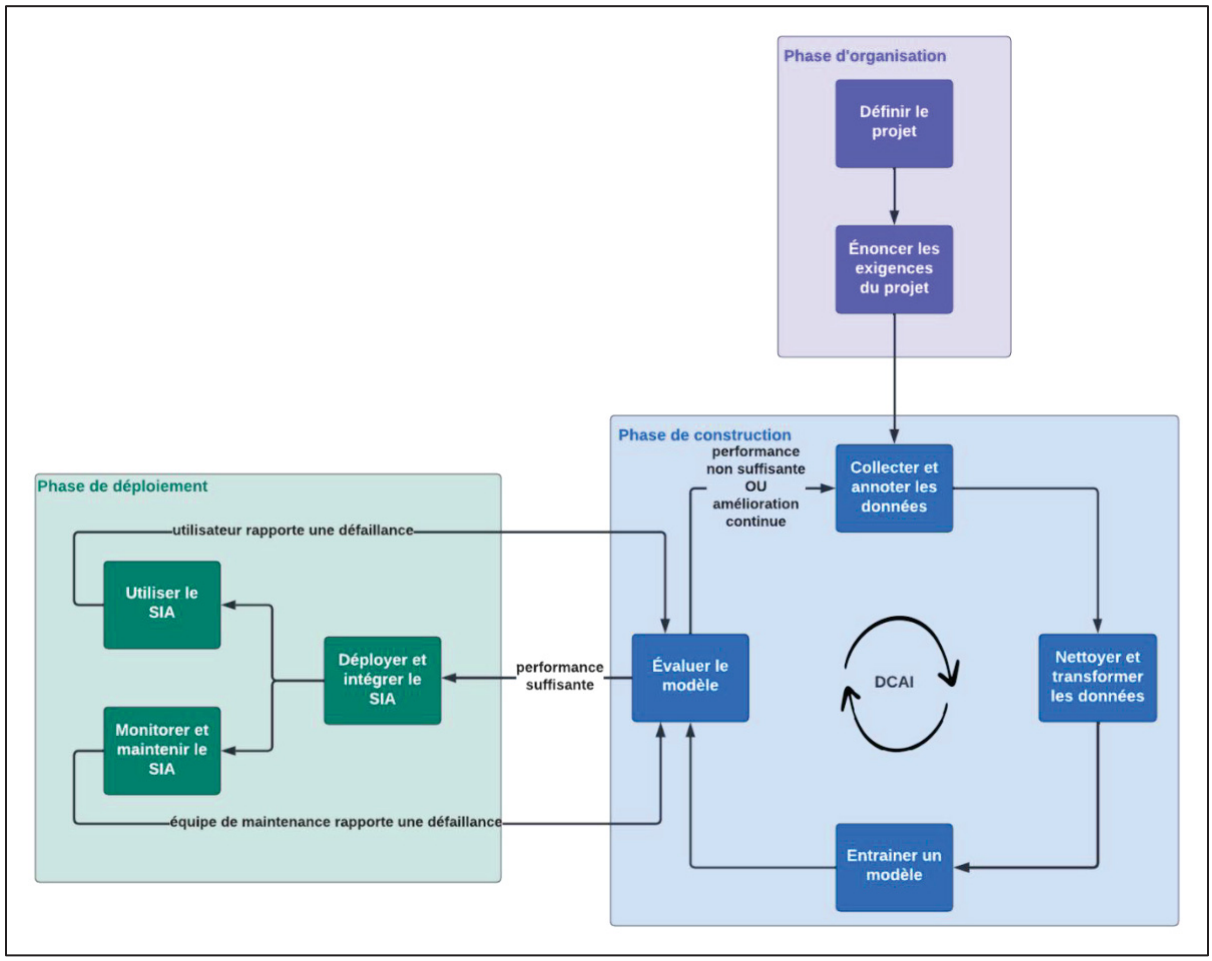


Figure 1.5 Cycle de vie d'un système d'intelligence artificielle incorporant l'approche Data-Centric AI

En revanche, l'approche DCAI contribue à l'explicabilité et à la transparence des systèmes d'IA en évitant une complexification inutile des modèles. L'approche DCAI privilégie la clarté et la simplicité en se concentrant sur la qualité des données. Cette approche réduit le besoin de recourir à des techniques algorithmiques excessivement complexes, qui peuvent souvent rendre les modèles d'IA opaques même pour leurs créateurs. Ainsi, des modélisations simples sont créées, sans sacrifier la performance.

1.6.3 Cycle de vie d'un système d'intelligence artificielle des petites et moyennes entreprises

Vitrine IA Québec signale que 91% des entreprises utilisant de l'IA, au Québec, sont des PME. En raison de la grande importance des PME dans l'économie, une transformation numérique à 2 vitesses, soit une beaucoup plus rapide pour les grandes entreprises que les PME, n'est pas envisageable. Les PME jouent un rôle crucial dans le développement économique et l'innovation, et leur intégration dans l'écosystème de l'IA est indispensable à la réussite économique québécoise. Les PME font face à diverses difficultés afin d'être en mesure d'implémenter des SIA et plus précisément des SIA de confiance. Cette section a pour objectif d'explorer l'intégration de l'approche DCAI dans le milieu des PME afin de valider son utilité.

Le DCAI se révèle donc être une approche précieuse pour les PME, adaptée à leurs besoins et contraintes spécifiques. Contrairement aux grandes entreprises qui disposent de vastes quantités de données et de ressources abondantes, les PME peuvent tirer un avantage significatif de l'approche DCAI qui met l'accent sur la qualité des données plutôt que sur leur volume. En valorisant la qualité des données plutôt que leur quantité, le DCAI permet une adhésion plus large à l'IA dans divers secteurs (Majeed et Hwang, 2023). Aussi, les PME peuvent être plus agiles et rapides pour adopter de nouvelles technologies, telles que l'IA, car elles ont généralement une structure organisationnelle plus simple. L'agilité et le développement itératif étant au cœur de l'approche DCAI, les PME possèdent un avantage concurrentiel certain pour la mise en place de ce processus.

Crockett *et al.* (2021) a identifié les difficultés vécues par les PME utilisant l'IA, grâce à des tables rondes menées avec 20 PME de divers secteurs. Les principales difficultés sont la disponibilité des ressources (personnes et temps), les compétences actuelles ainsi que les exigences de formation pour obtenir un emploi dans une PME. Ces obstacles ont un impact significatif sur le cycle de vie de développement d'un SIA notamment en limitant l'accès aux technologies avancées et en exigeant une approche plus agile et créative. Crockett *et al.* (2021) a défini les stades d'un cycle de vie d'une IA, au sein d'une PME. Les phases y sont moins détaillées que celles observées dans un cycle de développement d'une grande entreprise, du

fait d'une planification moins approfondie. Les ressources limitées forcent les PME à exploiter les concepts d'agilité et de créativité pour la planification d'un projet en IA.

Par exemple, les PME ont souvent des accès limités aux données. L'une des solutions qui s'offrent à elles est d'utiliser des données publiques ou d'envisager une collaboration afin d'obtenir des données provenant de sources externes. L'approche DCAI favorise de telles collaborations en offrant la possibilité d'ajouter des données au projet, de manière graduelle. Par manque de personnel, les PME envisagent fréquemment l'utilisation d'outils de développement préconçus pour faire la construction et l'évaluation de modèles. Cette méthode permet de développer rapidement une IA, à moindre coût et effort. Aussi l'approche DCAI amène à l'utilisation d'architecture simple, ce qui rend les connaissances plus accessibles aux employés des PME. Le déploiement est généralement plus facile pour les PME en raison de leur agilité. En revanche, les PME se doivent d'être plus créatives afin d'allouer des ressources dans la maintenance afin de maintenir les SIA performants.

En définitive, l'approche DCAI offre aux PME une voie viable pour intégrer efficacement l'IA dans leurs opérations, en surmontant les obstacles liés aux ressources et en maximisant l'utilisation des données disponibles, tout en favorisant l'agilité et l'innovation.

1.7 Parties prenantes d'un système d'intelligence artificielle explicable

Les parties prenantes, définies comme des individus ou des groupes qui peuvent influencer ou être influencés par les décisions et actions associées à un système, sont au centre de la mise en œuvre réussie d'un SIA explicable. Au travers des différents cadres proposés dans l'état de l'art, les besoins et les objectifs en matière d'explicabilité sont déterminés principalement par la catégorie à laquelle appartient un utilisateur. On peut entre autres nommer Arya et al. (2019), Langer et al. (2021) et Dhanorkar et al. (2021).

La difficulté d'identifier et de caractériser ces différentes parties prenantes et leurs besoins en matière d'explicabilité est un défi majeur. Tel qu'identifié par Suresh et al. (2021), la plupart

des recherches se concentrent sur l'une ou l'autre de ces méthodes pour classifier les acteurs : soit en se basant sur leur expertise (connaissances), soit en identifiant leur rôle par rapport au système intelligent. Au travers des différents cadres, les besoins et objectifs en matière d'interprétabilité sont principalement déterminés par la catégorie à laquelle appartient un utilisateur. Les prochaines sections se penchent sur l'implémentation de ces cadres ainsi que sur les avantages et difficultés engendrés par ceux-ci.

Si chacune de ces méthodologies présente des avantages, elles comportent également des lacunes. C'est dans cette optique que nous explorons une perspective novatrice qui cherche à fusionner les forces de ces deux approches. Dans les sections suivantes, nous nous pencherons sur la nécessité d'une telle fusion, ses implications et la manière dont elle pourrait redéfinir la façon dont nous envisageons l'interaction entre les utilisateurs et les SIA.

1.7.1 Caractériser les parties prenantes en fonction de leurs expertises

La caractérisation des utilisateurs ou des parties prenantes selon leurs expertises vis-à-vis des SIA représente un axe qui était majoritairement exploré dans la littérature, avant les années 2020. Cette caractérisation part du principe que les connaissances de la partie prenante dictent ses besoins en XIA.

L'état de l'art offre diverses approches à cette fin. Par exemple, Yu et Shi (2018) proposent une classification des parties prenantes basée sur leurs connaissances dans le domaine de l'IA. Ils distinguent quatre niveaux d'expertise : débutant, praticien, développeurs, et experts. Dans une autre étude, Mohseni et al. (2020) identifient trois catégories pertinentes pour la XIA : les novices en IA, les experts en IA et les experts en données. Aussi, Suresh et al. (2021) recommandent une division basée sur trois domaines spécifiques : le domaine du SIA, celui des données et le domaine environnemental.

Pour mieux comprendre ces domaines :

- Le domaine du SIA englobe les compétences nécessaires pour rechercher, développer, utiliser et déployer des modèles d'apprentissage automatique.
- Le domaine des données réfère aux compétences essentielles pour collecter, organiser, analyser et communiquer les données avec lesquelles le modèle est formé et prend ses décisions.
- Le domaine environnemental concerne la compréhension des environnements dans lesquels les interactions homme-IA ont lieu.

Cependant, plusieurs problématiques émergent de ces approches. Premièrement, elles ne tiennent pas compte de l'évolution des connaissances des utilisateurs au fil du temps. Imaginons un système de recommandation de films basé sur l'IA qui s'adapte aux préférences des utilisateurs. Quand un utilisateur s'inscrit pour la première fois, il est classé comme « débutant » en fonction des connaissances présumées de ce système. Sur la base de ce niveau d'expertise, le système lui propose une interface simplifiée avec des recommandations générales. Avec le temps, cet utilisateur interagit régulièrement avec le système, regardant de nombreux films et fournissant des évaluations. Malgré cela, le système, basé sur sa classification initiale de « débutant », ne lui offre pas les fonctionnalités avancées, comme la possibilité d'explorer des genres moins courants ou d'accéder à des analyses détaillées des films.

Deuxièmement, ces classifications, qui s'appuient essentiellement sur des notions cognitives d'expertise, négligent la riche connaissance tacite et l'expérience pratique des parties prenantes. Par exemple, une usine utilise un SIA afin d'analyser les données, faire des prédictions sur les performances des machines et fournir des explications sur les causes potentielles des pannes détectées. Afin de fournir des explications adéquates aux opérateurs et réparateurs des machines, ceux-ci sont classés en fonction de leurs formations ainsi que de leurs qualifications académiques. En revanche, l'un des opérateurs possède une expérience de 30 ans dans l'usine mais n'a pas de formation académique formelle dans le domaine. Sur la base des critères cognitifs, le système XIA le classe comme « intermédiaire » alors qu'il a développé une connaissance tacite profonde au fils des ans : il peut souvent détecter un

problème rien qu'en écoutant le bruit d'une machine ou en sentant une légère vibration. Sa connaissance est basée sur des années d'expérience pratique, mais elle est difficile à quantifier ou à classifier de manière traditionnelle.

Enfin, ces caractérisations ne distinguent pas suffisamment les parties prenantes ayant des niveaux d'expertise similaires, mais des objectifs différents. Par exemple, supposons qu'un hôpital utilise un système XIA pour aider les médecins à diagnostiquer des maladies rares. Deux médecins utilisent le système. Les deux sont classés comme « experts » en médecine. Le premier médecin est un chercheur médical qui utilise la XIA pour examiner les cas atypiques, étudier les nouvelles tendances et comprendre en profondeur les mécanismes des maladies. Les explications détaillées et techniques sont essentielles pour lui. Le deuxième médecin, bien qu'il soit aussi expert que le premier, utilise la XIA en consultation pour expliquer rapidement et clairement aux patients leurs diagnostics. Pour lui, une explication technique approfondie est moins utile qu'une explication simplifiée et visuelle qu'il peut montrer à son patient pour une meilleure compréhension. Leurs objectifs avec le système sont distincts.

1.7.2 Caractériser les parties prenantes en fonction de leurs rôles

La caractérisation des utilisateurs ou des parties prenantes selon leur rôle fonctionnel vis-à-vis des SIA représente un axe majeur de recherche dans la littérature. Cette caractérisation part du principe que le rôle d'une personne au sein d'une organisation (ou pendant une interaction humain-IA) influence directement ses besoins en matière d'explicabilité.

Hussain et al. (2021) ont mis en avant trois rôles clés pour le développement d'un SIA considéré comme critique, l'utilisation d'une automobile complètement autonome, à savoir :

- Ingénieurs et scientifiques : ceux qui conçoivent et développent les systèmes ;
- Éthiciens : experts qui évaluent l'impact éthique de l'IA ;
- Utilisateurs finaux et consommateurs : personnes pour lesquelles le système a été conçu.

Il est surprenant de constater que les entités de régulation ne sont pas considérées par ces auteurs comme une partie prenante essentielle. Ce choix, non justifié dans leur travail, pose question. Suresh *et al.* (2021) critiquent cette approche, soulignant que les cadres basés uniquement sur les rôles ne parviennent pas à segmenter l'espace du problème de manière suffisamment détaillée et modulaire. Ils citent, par exemple, le cas des modèles de diagnostics cliniques. Dans ce contexte, la catégorie des « consommateurs » pourrait englober autant les médecins que les patients. Or, ces deux groupes, malgré leur rôle similaire dans la consommation du modèle, auraient besoin d'explications différentes en raison de leurs niveaux d'expertise médicale distincts.

Reconnaissant ces limites, de récents travaux ont cherché à affiner la caractérisation des parties prenantes par rôle. Liao *et al.* (2021) ont notamment proposé un cadre intégrant de nouvelles catégories comme les « parties affectées par la décision ». Leur proposition segmente les parties prenantes comme suit :

- Développeurs de modèles : axés sur l'amélioration ou le débogage du modèle ;
- Propriétaires d'entreprises ou administrateurs : centrés sur l'évaluation de la capacité et la conformité réglementaire des applications IA ;
- Preneurs de décisions: utilisateurs directs d'applications de soutien à la décision basées sur l'IA ;
- Groupes Impactés : individus ou entités susceptibles d'être affectés par les décisions de l'IA ;

Des cadres similaires ont été suggérés par Arya *et al.* (2019), Langer *et al.* (2021) et Dhanorkar *et al.* (2021). Par exemple, Dhanorkar *et al.* (2021) différencient les scientifiques de données et les développeurs. Dans les propositions de Langer *et al.* (2021) et Liao *et al.* (2021), on retrouve le même regroupement de parties prenantes. Si les nuances entre ces cadres résident principalement dans la nomenclature et la précision de la segmentation des parties prenantes,

un point fondamental demeure : la distinction entre les parties prenantes affectées par la décision et les utilisateurs directs est largement reconnue dans la littérature récente.

Le développement des technologies de XIA s'est accéléré ces dernières années, créant ainsi un besoin urgent de standards et de cadres régissant leur conception et leur déploiement. Bien que la *Institute of Electrical and Electronics Engineers* (IEEE) n'ait pas encore établi un standard spécifique pour le développement de la XIA, il a franchi une étape significative en 2021 avec la publication du 7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design. Ce standard a comme objectif de permettre aux organisations de concevoir des systèmes en tenant compte explicitement des valeurs éthiques individuelles et sociétales, telles que la transparence, la durabilité, la confidentialité, l'équité et la responsabilité, ainsi que des valeurs généralement prises en compte dans l'ingénierie des systèmes, telles que l'efficacité et l'efficience (IEEE, 2021). Le tableau 1.3, proposé par le standard IEEE 7000-2021, met en évidence la structure de parties prenantes en fonction des rôles de chacune.

Une des particularités de ce cadre est l'introduction des « Opposants » en tant que parties prenantes. Cette catégorisation reconnaît l'importance de considérer ceux qui peuvent avoir des réserves, des objections, et qui veulent limiter la création ou l'utilisation adéquate d'un produit. L'IEEE (2021) introduit le fait que le niveau de transparence offert par les propriétaires de l'information est souvent en corrélation avec la confiance qu'ils ont en la capacité du participant à aligner ses propres valeurs avec les leurs, ainsi que la confiance dans la capacité du participant à protéger ses informations. Par exemple, alors que le propriétaire d'un modèle pourrait être enclin à divulguer le fonctionnement d'un SIA à un utilisateur, il pourrait en même temps être réticent à partager cette même information aux autorités externes et encore plus à un opposant. Le niveau de confiance de divulgation de l'information de la part des propriétaires serait aussi influencé par leur évaluation de la capacité du participant à comprendre et à appliquer les détails techniques mis à leur disposition.

Tableau 1.3 Parties prenantes d'un système d'intelligence artificielle,
Tiré de Standard IEEE 7000-2021

Partie prenante	**Exemple(s)**	**Intentions**
Internes	Propriétaires du projet, haute direction de l'organisation, architectes et concepteurs du système, analystes, gestionnaires, ingénieurs système, ingénieurs logiciels, etc.	Les parties prenantes internes peuvent être moins affectées par les domaines où des préoccupations éthiques surgissent, car elles ne sont souvent pas les utilisateurs du système.
Utilisateurs	Opérateurs du système mais aussi les personnes qui bénéficient de l'utilisation du système ou en sont affectés négativement. Également celles dont les données personnelles sont stockées dans un système.	Obtenir un service ou un produit qui répond à leurs besoins. Les besoins varient en fonction du service ou du produit ainsi que de l'utilisateur qui en fait l'acquisition.
Opposants	Les concurrents, les pirates informatiques ou les opposants à l'organisation, produits ou services développés.	Intentions néfastes diverses. Celles-ci varient en fonction de l'opposant.
Autorités externes	Les régulateurs gouvernementaux et les groupes de défense externes, les évaluateurs tiers, les courtiers de données et les entrepreneurs indépendants de vérification et de validation (IV&V).	Les normes culturelles et les valeurs éthiques peuvent différer de celles du propriétaire du système, peuvent entraîner des conflits de valeurs et restreindre les décisions du propriétaire du système. Elles peuvent mettre en évidence des défauts ou des hypothèses non déclarées qui ont faussé les choix éthiques de l'organisation.

L'une des critiques fréquemment observées lors de l'utilisation d'un cadre fondé sur les rôles est que l'expertise au sein même des catégories de parties prenantes est souvent négligée (Suresh *et al.*, 2021). Un cadre de parties prenantes qui introduit « l'équipe de développement » ne tiendrait pas en compte des différences notables entre le rôle de scientifique de données et ingénieur de données, par exemple. Plusieurs éléments diffèrent entre ces deux parties prenantes comme notamment les tâches en lien avec le SIA et le moment d'interaction avec le SIA dans les phases de développement. On peut donc pressentir que leurs besoins en

explicabilité et transparence varient. Il a été noté que plusieurs de ces cadres ont la volonté de présenter des parties prenantes générales afin que celles-ci puissent être utilisées pour tous les cas d'usages. Cependant, en ce qui nous concerne, nous émettons la conclusion que cela n'est pas une approche à prioriser dû à la perte d'information pertinente. Il serait donc judicieux d'établir un cadre détaillé, plutôt que généraliste, afin de distinguer les nuances entre les parties prenantes et de mieux identifier leurs besoins. En fonction des besoins du projet, des parties prenantes pourraient être supprimées de la liste et ainsi l'on conserverait une quantité d'informations plus grande, nécessaires à l'identification des besoins en XIA.

Aussi, il est noté que même au sein d'un même rôle, l'expertise peut varier. Piorkowski *et al.* (2021) ont relevé cette disparité, soulignant que l'expertise diversifiée des membres d'une même équipe, au sein d'un même rôle, provient souvent de leurs formations et expériences variées. Un cadre créé en fonction des rôles ne permet pas de faire ces distinctions. Certes la caractérisation des parties prenantes selon leur rôle offre une première approximation utile des besoins en matière d'explicabilité des SIA, plus précisément concernant leurs interactions avec le SIA. Néanmoins, elle est insuffisante. Une segmentation plus nuancée, tenant compte à la fois du rôle et du niveau d'expertise des différentes parties prenantes, semble nécessaire pour élaborer un cadre adapté aux besoins réels en matière d'explicabilité de l'IA.

1.7.3 Rôles et expertises : une perspective combinée pour la caractérisation des parties prenantes

Il est largement reconnu que l'identification des parties prenantes d'un système d'IA (SIA) est capitale pour une meilleure compréhension et identification des besoins en explications de l'IA (XIA). Toutefois, le domaine n'a pas encore atteint un consensus concernant un cadre précis pour la caractérisation des parties prenantes. En effet, deux concepts majeurs émergent : caractériser les parties prenantes en fonction de leurs expertises OU caractériser les parties prenantes en fonction de leurs rôles. Il est essentiel de souligner l'importance de ce OU, car il suggère un choix entre deux paradigmes distincts.

Au sein même de ces deux concepts, aucun cadre n'est unanimement accepté par la communauté scientifique. Notre analyse de la littérature a montré que chacune de ces approches présente ses propres limites. Caractériser les parties prenantes selon leurs expertises ne tient pas compte de leurs intentions, tandis que la caractérisation basée sur le rôle néglige la diversité des connaissances au sein d'un même groupe de rôles.

Néanmoins, un constat intéressant se dégage : les défauts associés à un cadre sont souvent atténués ou corrigés par les atouts de l'autre. Cela suggère une perspective novatrice pour l'avenir. Plutôt que de choisir l'un ou l'autre, il serait plus judicieux d'envisager l'utilisation complémentaire de ces deux cadres.

- Caractériser les parties prenantes par leurs rôles : Cette optique permet de saisir les intentions des parties prenantes ainsi que leurs interactions avec le SIA. Cela offre la possibilité de déterminer quelles sont les informations requises pour chaque partie prenante et à quel moment précis, dans le cycle de vie et d'utilisation du système, elles sont requises.
- Caractériser les parties prenantes par leurs expertises : Cette vision se concentre sur la manière dont l'information est présentée aux différentes parties prenantes. Elle permet d'adapter la présentation des informations en fonction des parties prenantes afin de garantir une assimilation rapide et efficace. Cette façon de procéder tient compte des compétences et des connaissances de chaque groupe.

En combinant ces deux approches, nous pourrions concevoir un cadre qui reconnaît à la fois l'importance des rôles des parties prenantes et la nécessité d'adapter l'information selon leurs expertises. Cette fusion pourrait mener à une meilleure personnalisation des systèmes d'IA explicables, optimisant ainsi leur utilité et leur pertinence pour divers utilisateurs. Afin d'assurer une caractérisation des parties prenantes par leurs rôles pour faciliter l'identification des besoins en IA, il est aussi recommandé de définir des rôles de manière la plus granulaire possible, quitte à en supprimer lorsque ceux-ci ne s'appliquent pas à un projet. Aussi pour augmenter l'efficacité de la caractérisation des parties prenantes par leurs expertises, il est

important de définir non seulement les expertises elles-mêmes, mais aussi le degré d'expertise puisque celui-ci peut varier (i.e. novice, intermédiaire, expert).

CHAPITRE 2

MÉTHODOLOGIE

La revue de littérature a permis d'identifier plusieurs lacunes significatives dans le domaine de l'intelligence artificielle explicable (XIA), en particulier dans le contexte québécois.

Premièrement, des lacunes dans la clarification des concepts et l'utilisation de termes vagues dans la législation québécoise portant sur la XIA (voir section 1.4). Ces lacunes génèrent de la confusion, notamment pour les ingénieurs spécialisés en IA souhaitant se conformer aux normes, sans vraiment comprendre nettement les attentes. Cette ambivalence crée un terrain propice à des utilisations incorrectes de la XIA, et surtout, limite la création de requis fonctionnels qui prennent en considération les mentions légales.

Deuxièmement, l'absence de méthodologie pour développer des mesures concrètes relatives aux engagements éthiques (voir section 1.5). En permettant une compréhension approfondie du processus décisionnel des SIA, la XIA, en tant que principe intermédiaire établi par les entreprises, peut contribuer au respect des faits moraux fondamentaux, tel que la transparence. Pour atteindre cet objectif, les entreprises doivent développer des mesures concrètes, et être capables de valider si leur mise en œuvre répondent aux attentes éthiques de la XIA. Actuellement, une telle méthodologie ne fait pas encore l'objet d'une large adoption dans la littérature spécialisée sur la XIA. Il est donc complexe pour les ingénieurs IA de définir des requis fonctionnels qui intègrent des mesures concrètes de la XIA pour assurer le respect des engagements éthique de leur employeur.

Troisièmement, le manque d'uniformité dans la définition du cycle de vie des SIA (voir section 1.6). Les entreprises qui créent des SIA décrivent le cycle de vie de ces systèmes de différentes manières (Ng, 2021). Cette définition est importante car elle contribue à la clarification du contexte d'utilisation de l'IA, un élément clé permettant de définir des requis fonctionnels. De plus, le choix du cycle de vie de l'IA peut avoir un impact sur la transparence de celle-ci. Notre

revue de littérature identifie que le choix d'un cycle de vie *Data-Centric* aurait cet effet sur le système.

Quatrièmement, bien qu'il existe une diversité de métriques permettant d'évaluer la robustesse des explications générées par l'intelligence artificielle explicable (XIA), il n'existe à ce jour aucune directive claire sur le choix de la métrique à utiliser selon le contexte (voir section 1.7). De nombreuses métriques ont été proposées au sein de la communauté scientifique, mais aucune standardisation n'a été établie. Il n'y a pas de consensus sur celles qui devraient être priorisées ni sur leur pertinence selon les cas d'usage. En conséquence, les chercheurs et praticiens emploient souvent différentes métriques sans justification rigoureuse. À ce jour, aucune méthodologie ne permet d'identifier de manière structurée les besoins spécifiques en XIA, ni de guider la sélection des métriques les plus appropriées pour répondre à ses besoins spécifiques.

Cinquièmement, des lacunes dans l'identification des parties prenantes (voir section 1.8). Il est largement reconnu que l'identification des parties prenantes d'un système d'IA est cruciale pour une meilleure compréhension et identification des besoins en XIA. Toutefois, il n'y a pas encore de consensus à propos d'un cadre précis pour la caractérisation des parties prenantes. Deux tactiques majeures émergent : caractériser les parties prenantes en fonction de leurs expertises OU caractériser les parties prenantes en fonction de leurs rôles. Notre revue de littérature nous permet de conclure que, plutôt que de privilégier l'une au détriment de l'autre, il faudrait prioriser une approche complémentaire de ses deux tactiques.

Face à ces constats, la présente recherche, s'appuyant sur une approche scientifique et des principes d'ingénierie, vise à réduire l'écart entre les explications générées par les systèmes XIA et les besoins réels des parties prenantes, en proposant une méthodologie qui permet d'identifier les besoins, exigences et contraintes en matière de XIA. Afin d'atteindre cet objectif, nous décomposons le projet en cinq phases principales.

Premièrement, la traduction des exigences normatives et éthiques en éléments directement mobilisables par les ingénieurs IA. Nous débutons par une analyse des textes législatifs québécois pertinents en matière d'intelligence artificielle, notamment les récentes modifications introduites par la Loi 25. Cette analyse vise à identifier les obligations explicites et implicites relatives à la XIA, ainsi que les zones grises susceptibles d'interprétation. Sur cette base, des recommandations pratiques sont formulées afin de rendre ces exigences opérationnelles dans les contextes de développement et de déploiement de systèmes IA. Ces travaux sont détaillés dans l'article « Regard de l'ingénierie sur les lois québécoises en matière d'intelligence artificielle explicable (XIA) » (voir chapitre 3). En parallèle, les grands principes éthiques portés par les référentiels internationaux sont traduits en indicateurs concrets. L'enjeu est ici de proposer une méthode permettant aux professionnels de transformer des engagements éthiques généraux (comme la transparence et la XIA) en exigences fonctionnelles et mesurables, intégrées dès les premières étapes du développement. Ces éléments font l'objet d'une exploration dans l'article « Le développement d'une IA explicable : entre principes éthiques et mesures concrètes » (voir chapitre 4).

Deuxièmement, la conception des protocoles et cadres de références nécessaires à une compréhension fine des besoins en XIA par les parties prenantes du système. Ce volet est particulièrement développé dans l'article Merging Roles and Expertise: Redefining Stakeholder Characterization in Explainable Artificial Intelligence (voir chapitre 5), qui propose un protocole d'identification des parties prenantes reposant sur une double caractérisation : d'une part, leur rôle au sein du cycle de vie d'un système d'IA (utilisateur, développeur, décideur, régulateur, etc.), et d'autre part, leur niveau d'expertise en intelligence artificielle. Cette approche combinée permet d'élaborer une cartographie des parties prenantes qui tient compte à la fois de leur interaction avec le système et de leur degré de vulgarisation nécessaire face aux décisions automatisées. L'article introduit également une méthode de priorisation des besoins en XIA, fondée sur la hiérarchisation des parties prenantes selon ces critères.

Troisièmement, le développement d'une méthodologie structurée pour formaliser les besoins, exigences et contraintes en matière de la XIA. Cette méthodologie, désignée sous le nom de Explainable AI Requirements Specification, est détaillée dans l'article « Évaluation structurée des besoins, exigences et contraintes en intelligence artificielle explicable avec l'outil XAIRS » (voir chapitre 6). Elle s'appuie sur les fondements établis dans les chapitres précédents, en intégrant les apports relatifs à la traduction des exigences légales québécoises (chapitre 3), à l'opérationnalisation des principes éthiques (chapitre 4), ainsi qu'à l'identification et à la priorisation des parties prenantes (chapitre 5), permettant de guider les équipes de développement dans la définition rigoureuse des requis en matière de XIA, en tenant compte des spécificités contextuelles, réglementaires et humaines propres à chaque projet.

Quatrièmement, mettre à l'épreuve la méthodologie Explainable AI Requirements Specification à travers une évaluation empirique et analyser les données recueillies. Cette démarche est aussi décrite dans le chapitre 6. Cette évaluation prend la forme d'une étude de cas appliquée à un projet industriel réel, issu d'un secteur critique où la XIA constitue une exigence essentielle (par exemple, la santé, la finance ou la justice). Un acteur clé du projet tel qu'un responsable technique ou un chef de projet IA, est sélectionné pour participer à une enquête visant à évaluer la pertinence, l'utilité et la faisabilité des outils et protocoles proposés. Un questionnaire est conçu à cet effet (voir Annexe III), combinant des questions fermées (sur échelles de Likert) et des questions ouvertes. Ce questionnaire permet d'évaluer dans quelle mesure les contributions proposées répondent aux besoins concrets des professionnels, facilitent leur travail de conception, et contribuent à une meilleure conformité juridique et éthique. Il vise également à recueillir des commentaires qualitatifs permettant d'identifier d'éventuelles améliorations ou ajustements à apporter à la méthodologie.

Cinquièmement, la validation des apports de la recherche, présentée au chapitre 7. Cette phase vise à confronter l'ensemble du travail accompli aux objectifs initiaux, afin d'évaluer dans quelle mesure ces derniers ont été atteints. Elle offre un moment de synthèse critique permettant de mesurer la pertinence, la portée et les limites des contributions proposées dans le cadre de cette démarche scientifique.

CHAPITRE 3

REGARD DE L'INGÉNIERIE SUR LES LOIS QUÉBÉCOISES EN MATIÈRE D'INTELLIGENCE ARTIFICIELLE EXPLICABLE (XIA)

Camélia Raymond[a], Sylvie Ratté[b], Marc-Kevin Daoust[c]

[a, b, c] Département de Génie Logiciel et des TI, École de Technologie Supérieure
Département des Enseignements Généraux, École de Technologie Supérieure
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article soumis pour publication, Lex Electronica, avril 2026

3.1 Résumé

L'article 12.1 de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels (LPRPSP) aborde de manière implicite l'eXplicabilité de l'Intelligence Artificielle (XIA) en exigeant des entreprises qu'elles fournissent aux personnes concernées des informations sur les données personnelles utilisées pour prendre des décisions automatisées, ainsi que sur les raisons et les principaux facteurs à l'origine de ces décisions. Une analyse de cette disposition révèle des lacunes et des incohérences entravant la mise en œuvre ingénieuse de ses exigences. Les interactions complexes entre l'ingénierie et le droit ainsi que les interactions qui influencent la mise en œuvre de l'IA dans un cadre juridique évolutif sont analysées. L'article explore les difficultés posées par les exigences de la *LPRPSP* concernant la XIA, en proposant des solutions pour une meilleure harmonisation entre les principes d'ingénierie et les exigences légales.

3.2 Summary

Article 12.1 of the Act to modernize legislative provisions as regards the protection of personal information (LPRPSP) implicitly addresses eXplainable Artificial Intelligence (XAI) by requiring companies to provide individuals with information about the personal data used to make automated decisions, as well as the reasons and main factors behind these decisions. An

analysis of this provision reveals gaps and inconsistencies that hinder the ingenious implementation of its requirements. The complex interactions between engineering and law and how these interactions influence the implementation of AI within an evolving legal framework are examined. The article explores the challenges posed by the LPRPSP's requirements related to XIA and proposes solutions for better harmonization between engineering principles and legal requirements.

3.3 Introduction

La Loi 25 (Gouvernement du Québec, 2021c), une législation ambitieuse et innovante en matière de protection des renseignements personnels, a suscité un vif intérêt dans le paysage médiatique québécois. Cette loi, officiellement intitulée *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*, a été saluée dans ses premières apparitions médiatiques avec des titres élogieux. Des formulations telles que « Le Québec aux avant-postes en matière de protection des données personnelles » (Nadeau, 2022) et « La protection de la vie privée au Québec a désormais beaucoup de mordant » (McKenna, 2021) illustraient l'enthousiasme initial. Néanmoins, depuis la période d'entrée en vigueur progressive, le ton a changé dans les articles. Des titres tels que « Québec respecte mal sa propre loi sur les renseignements personnels » (McKenna, 2023) ou encore « Nouvelles règles sur les renseignements personnels : Des PME à la traîne » (Joncas, 2023) dénotent moins d'enthousiasme. Depuis son entrée en vigueur, la Loi 25 a captivé l'attention du public, engendrant une participation et un engagement remarquables dans son application.

Au-delà de son adoption médiatique, la Loi 25 établit un cadre législatif clé pour la protection des données personnelles, particulièrement pertinent dans le contexte de l'intelligence artificielle (IA). En imposant des normes de gestion et de sécurité des données, elle soulève des enjeux fondamentaux de confiance envers ces technologies. Son impact dépasse les cercles techniques, alimentant un débat public élargi où la société québécoise joue un rôle de premier plan.

Dans ce contexte, « l'IA de confiance » a été proposée par la communauté scientifique et d'ingénierie comme étant une solution nécessaire à la création et à l'implémentation d'IA répondant aux craintes du public. Mais concrètement, que représente l'idée d'une intelligence artificielle de confiance ? L'IA de confiance est un idéal difficile à définir. Malgré l'utilisation fréquente de l'expression « IA de confiance », cette notion n'a pas encore été explicitement déterminée dans le paysage juridique québécois. D'un point de vue Européen, il est établi qu'une IA de confiance s'obtient par le respect des sept principes suivants : le contrôle humain, la robustesse et la sécurité, la conformité et la gouvernance, l'explicabilité et la traçabilité, la diversité et l'équité, l'impact social et environnemental, et les responsabilités associées au traitement (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

Par ailleurs, le concept d'une IA explicable (XIA - eXplainable Artificial Intelligence) est abordé dans la Loi 25. Une XIA peut être perçue comme un morceau du casse-tête que représente l'IA de confiance. Une XIA permet, conjointement avec d'autres principes, de répondre au manque de confiance en l'IA. Selon Les lignes directrices en matière d'éthique pour une IA digne de confiance, le concept d'explicabilité, dans le domaine de l'IA, se définit comme suit :

« L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'intelligence artificielle et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'intelligence artificielle). L'explicabilité technique suppose que les décisions prises par un système d'intelligence artificielle peuvent être comprises et retracées par des êtres humains. [...] Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur) » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

Au Québec, les systèmes d'IA explicable doivent répondre à des contraintes juridiques. Cet article se penche, à la section 3.4, sur les lois québécoises qui traitent de la XIA et sur les défis

auxquels font face les ingénieurs pour mettre en place une IA de confiance. L'ingénierie de l'IA, désireuse de se conformer aux lois québécoises sur la question, se heurte au manque de clarté dans la terminologie, ainsi qu'aux limitations technologiques dans le domaine de la XIA, éléments explorés dans la section 3.5 de cet article. Finalement, notre étude de la Loi modernisant des dispositions législatives en matière de protection des renseignements personnels, présentée à la section 3.6, permet d'émettre des recommandations en direction des entités de régulations québécoises, des ingénieurs IA, des organismes publics ainsi qu'a l'intention de la communauté scientifique.

Cet article propose une analyse critique de la Loi 25 à la lumière des enjeux liés à l'intelligence artificielle explicable (XIA). Il met en évidence les tensions entre les exigences juridiques et les capacités technologiques actuelles, tout en offrant des pistes concrètes pour favoriser une meilleure harmonisation entre le droit et l'ingénierie. L'objectif est de contribuer à une mise en œuvre plus cohérente, compréhensible et techniquement réaliste des obligations liées à l'explicabilité, en particulier dans le contexte de décisions entièrement automatisées.

3.4 Législation québécoise des systèmes d'intelligence artificielle explicables

La récente législation québécoise en matière de protection des renseignements personnels, la Loi 25, traverse une phase critique de mise en œuvre. Ayant pour objectif de réglementer les domaines émergents utilisant des données personnelles, tels que les systèmes d'intelligence artificielle explicables, cette loi est actuellement source de débats. Selon le Rapport du sondage sur la Loi 25 (2023), parmi les entreprises ayant participé au sondage, 69% ont fait part de leur préoccupation quant au manque de précision des dispositions pratiques de cette loi, appelant à une définition plus claire et explicite des termes et des exigences. La conclusion de ce rapport est que la Loi 25 manque encore de clarté : elle suscite des inquiétudes et de la confusion (Gowling WLG & IAB Canada, 2023).

La transparence renvoie à la nécessité d'être ouvert et honnête sur les processus, les décisions et les actions. Elle est fondamentale, car elle touche à des valeurs morales essentielles comme

la confiance, la justice et l'intégrité. La XIA est alors définie comme un principe qui permet d'appliquer la transparence à des cas pratiques.

La *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*, entrée graduellement en vigueur au cours des années 2021 à 2024, a comme objectif d'offrir aux citoyens un meilleur contrôle de leurs renseignements personnels (Gouvernement du Québec, 2022). Elle modernise le cadre législatif afin de l'adapter à la réalité technologique d'aujourd'hui. Le traitement automatisé des renseignements personnels (par exemple, par les technologies de l'IA) est prévu par cette loi, car il s'agit là d'une réalité technologique qui est utilisée par plusieurs entreprises œuvrant au Québec. Notre analyse de cette loi a révélé que cette dernière aborde implicitement la XIA dans l'article 12.1 de la *Loi sur la protection des renseignements personnels dans le secteur privé* (LPRPSP). Par ailleurs, cette même disposition a été identifiée dans le Rapport du sondage sur la Loi 25 (2023) comme l'une des sections particulièrement sujettes à interprétation, engendrant ainsi des incertitudes.

« 12.1. Toute personne qui exploite une entreprise et qui utilise des renseignements personnels afin que soit rendue une décision fondée exclusivement sur un traitement automatisé de ceux-ci doit en informer la personne concernée au plus tard au moment où elle l'informe de cette décision.
Elle doit aussi, à la demande de la personne concernée, l'informer :
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.
Il doit être donné à la personne concernée l'occasion de présenter ses observations à un membre du personnel de l'entreprise en mesure de réviser la décision » (Gouvernement du Québec, 2021a).

Selon l'article 12.1 al. 3 de la LPRPSP, il est établi que toute personne a le droit de demander la rectification de ses renseignements personnels s'ils sont jugés incorrects ou obsolètes. Cette disposition pourrait influencer significativement les systèmes d'IA, notamment si ces

renseignements rectifiés ont été utilisés lors de leur phase d'apprentissage. La correction de ces données peut impacter les prédictions du système, y compris pour des cas ne concernant pas directement l'individu à l'origine de la demande de rectification. L'exigence imposée aux entreprises de modifier les données intégrées dans les algorithmes présente le risque d'altérer la performance et la robustesse de ces systèmes pour l'ensemble des utilisateurs. Cette situation met en lumière le défi de concilier le droit individuel à la correction des données personnelles avec la nécessité de préserver la fonctionnalité et l'efficacité des systèmes d'IA pour tous. Bien que cette section de la loi pourrait bénéficier de certaines nuances (par exemple, concernant la phase d'apprentissage des modèles), elle demeure clairement formulée et en adéquation avec les réalités technologiques.

L'article 12.1 al. 2 de la LPRPSP introduit une nouvelle responsabilité pour les entreprises québécoises. Il suggère, selon certaines interprétations, l'adoption du concept de « droit à l'explication », un principe couramment mentionné en Europe dans les débats sur l'IA explicable. Si cette interprétation est correcte, les entreprises du Québec devront expliquer à leurs clients les raisons des décisions prises par des systèmes complètement automatisés, c'est-à-dire sans intervention humaine dans le processus décisionnel. Bien que l'intention derrière cette exigence soit claire, la formulation utilisée dans la loi laisse une marge d'incertitude significative.

3.5 Quand les failles de l'ingénierie se heurtent à l'incohérence judiciaire

Nos analyses de l'article 12.1 de la LPRPSP, ont permis d'identifier des notions faisant défaut ainsi que des incohérences, d'un point de vue d'ingénierie. Les sections suivantes explorent ces notions du point de vue des experts en IA et les difficultés de mise en place de mesures concrètes en réponses aux interprétations possibles de l'article 12.1 de la Loi 25.

L'utilisation d'un vocabulaire imprécis tel que « des raisons, ainsi que des principaux facteurs et paramètres » laisse une grande place à l'interprétation. Puisque la XIA est un concept émergent, il n'y a pas encore de lignes directrices clairement établies et acceptées dans la

communauté scientifique (Tobey, 2019). Tant les entités de régulations que les spécialistes en ingénierie de l'IA se trouvent confrontés à une zone d'incertitude. Les sections 2.a à 2.c se consacrent à l'exploration de ces incertitudes. Elles visent à définir, d'un point de vue d'ingénierie, la signification précise des termes « raisons » ainsi que des « principaux facteurs » et « paramètres ». Elles visent aussi à déterminer les attentes pratiques et légales qui pourraient découler de ces définitions.

De plus, les demandes relatives à l'article 12.1 de la Loi 25 peuvent être perçues comme excessives pour certains systèmes d'IA, particulièrement ceux ayant un impact négligeable dans des applications peu critiques. L'absence de distinction sur le degré d'impact d'un système d'IA soulève des préoccupations qui sont abordées dans la section 2.d).

3.5.1 Définition du terme « raisons » dans le contexte de l'intelligence artificielle

Dans le domaine d'expertise de l'IA, le terme « raisons » fait référence à la logique apprise par l'algorithme pour émettre une prédiction. Une explication peut être locale ou globale. Une explication locale signifie que l'on motive le raisonnement pour une seule ou un nombre restreint de prédictions (Chamola et al., 2023). L'explication sera donc différente, en fonction de la personne qui en fait la demande. Une explication globale émet le raisonnement logique pour l'ensemble du traitement automatisé, c'est-à-dire la logique qui aurait été apprise, de manière générale, par l'IA (Chamola et al., 2023). Dans ce cas, toutes les personnes faisant la demande du raisonnement recevraient la même réponse qui ne serait pas personnalisée à la situation de chacun. Il est plus simple pour une entreprise de fournir une explication globale, mais est-ce que cela est réellement conforme à l'esprit de l'article 12.1 al. 2 de la LPRPSP ? Il est logique de penser qu'une explication locale serait plus appropriée puisque cela permet à un client de comprendre l'impact de l'algorithme sur son contexte précis.

Aussi, en comparant avec les lois européennes, un terme important est absent de l'article 12.1 al. 2 : « compréhensible ». Au Québec, un total de 44% des citoyens pensent savoir ce qu'est l'IA et 10% ne savent pas du tout ce que c'est. Seuls 13% affirment savoir précisément de quoi

il s'agit (Mila, 2020). Le terme « compréhensible » assure au lecteur de l'explication que celle-ci prendra en considération le niveau de connaissances en IA que l'on observe dans la population québécoise. Lorsque l'on pousse cette réflexion à l'extrême, l'absence de ce terme pourrait permettre aux entreprises de référencer, par exemple, un cours universitaire sur les réseaux de neurones afin de décrire les raisons pour lesquelles la prédiction a été émise. Cela ne serait évidemment pas adapté à la population québécoise.

Finalement, l'obligation de fournir des explications, justifiant le pourquoi d'une prédiction, est une tâche complexe pour les entreprises du fait des limites technologiques. La XIA est une sphère de recherche très active dont les solutions ne sont pas encore clairement définies, particulièrement pour des algorithmes complexes qui sont généralement les plus performants, tels que les réseaux de neurones (Chamola et al., 2023). Il est important de souligner que les techniques de XIA existantes ne garantissent pas nécessairement de fournir les véritables explications (Chamola et al., 2023). Un exemple typique d'explications est de fournir les caractéristiques d'entrées qui influencent le plus la prédiction. Par exemple, si l'on souhaite prédire le succès d'un(e) étudiant(e) à un examen, l'IA pourrait indiquer que les caractéristiques qui ont le plus influencé la prédiction sont : le niveau de stress de l'étudiant, le temps d'étude et les habitudes de sommeil. Si l'algorithme utilisé est de type « boîte noire », il n'est pas garanti que ses caractéristiques soient réellement celles qui impactent le plus la prédiction, alors que si l'algorithme est qualifié de « boîte en verre » ou de « boîte blanche », il propose un niveau de garanti plus élevé. Ce qui est inquiétant, c'est que l'information sur le type d'algorithme utilisé n'est généralement pas divulguée. En conséquence, les utilisateurs pourraient être convaincus qu'une explication de la logique est la bonne et ne pas être amenés à la remettre en question.

En prenant en considération la définition du terme « raison » dans le domaine de l'IA, une disparité entre la demande juridique et les capacités technologiques pourrait signifier que les entreprises doivent limiter leurs utilisations d'algorithmes complexes. Cela aurait bien sûr un impact majeur sur le marché québécois, en diminuant les performances des systèmes d'IA en industrie.

3.5.2 Définition des termes « principaux facteurs » dans le contexte de l'IA

Dans le domaine d'expertise de l'IA, les termes « principaux facteurs » possèdent diverses définitions. La définition de ces termes dépend du contexte de leur utilisation. Les différentes définitions possibles incluent notamment, les caractéristiques importantes, la définition du comportement de l'IA ainsi que les éléments qui influencent l'efficacité de l'IA.

La première définition serait celle des caractéristiques importantes. Les caractéristiques les plus importantes d'un ensemble de données utilisées pour entraîner une IA sont définies comme étant des facteurs. Ces caractéristiques, souvent appelées « features » en anglais, sont utilisées en entrée de l'IA pour prédire une sortie ou une cible. Par exemple, dans un modèle de prédiction de prix de l'immobilier, les caractéristiques pourraient inclure la taille de la maison et le nombre de chambres à coucher. Ces caractéristiques peuvent subir des transformations. Par exemple, au lieu de prendre en entrée de l'algorithme le nombre de chambres de la maison qui serait 4, l'algorithme prendrait uniquement une valeur binaire indiquant si le nombre de chambres de la maison est inférieur ou supérieur à 3. L'article 12.1 al. 2 de la LPRPSP ne mentionne pas si les principaux facteurs doivent être identifiés avant leurs transformations. Aussi, cette définition se rapproche de la demande de l'article 12.1 al. 1 qui exige de fournir les données en entrée à la demande de la personne concernée. Il est donc justifié de supposer que cette définition n'est probablement pas celle recherchée par le législateur de l'article 12.1 al. 2

La prochaine interprétation, dans une perspective d'ingénierie, serait les hyperparamètres. Les hyperparamètres sont des valeurs qui sont passées à une fonction, un programme ou un algorithme pour lui indiquer comment il doit s'exécuter. Ceux-ci sont définis par le développeur. Ils peuvent être utilisés pour spécifier des options de configuration, des données d'entrée ou des valeurs de retour. Selon le même principe, un programme de traitement de texte peut avoir un hyperparamètre qui indique la taille de la police à utiliser, un algorithme de tri peut avoir un hyperparamètre qui indique la manière dont les éléments doivent être triés, et

une fonction mathématique peut avoir un hyperparamètre qui indique une valeur initiale à utiliser dans le calcul.

Lors de l'entraînement d'une IA, les hyperparamètres servent à optimiser mathématiquement l'algorithme et à l'adapter aux données qui lui sont fournies. L'un des hyperparamètres les plus communs aux algorithmes d'IA est le taux d'apprentissage. Celui-ci doit être optimisé afin de permettre à l'algorithme d'atteindre de meilleures performances. Ces hyperparamètres sont en relation avec des notions mathématiques très poussées et complexes. Il n'est pas rare même qu'un ingénieur IA ne connaisse pas la signification ou l'impact de tous les hyperparamètres d'un algorithme d'IA. Plusieurs algorithmes possèdent plus d'une cinquantaine d'hyperparamètres. On peut citer, comme exemple, le modèle XGBoost (eXtreme Gradient Boosting) (XGBoost Developers, 2022), un algorithme couramment utilisé lors de compétitions en apprentissage automatique sur des données au format tabulaire. En comparaison, GPT-3, un algorithme beaucoup plus complexe, pourrait posséder plus d'une centaine d'hyperparamètres. Les détails techniques sur les hyperparamètres spécifiques et leur nombre sont généralement considérés comme des informations propriétaires par OpenAI, l'organisation derrière le développement de GPT-3. Les publications officielles d'OpenAI et les documents de recherche se concentrent principalement sur la structure du modèle, le nombre de paramètres (différent du nombre d'hyperparamètres), et les performances à travers différentes tâches, plutôt que sur les détails des hyperparamètres. En revanche, lors des débats de la Commission des institutions, le 27 mai 2021, Gaétan Barrette[1] a soulevé que cette loi (article 12.1 de la LPRPSP) ne devait pas forcer les entreprises à divulguer des informations dont elles seraient seules propriétaires, ce qui entraînerait un bris de confidentialité ou une entorse à la propriété intellectuelle (Assemblée nationale du Québec, 2019). Des exemples tels que GPT-3 montrent que les entreprises gardent souvent les hyperparamètres confidentiels. Ainsi, les « principaux facteurs » ne devraient pas être considérés comme les hyperparamètres de l'algorithme.

[1] Député de La Pinière pour le Parti Libéral du Québec, membre du Bureau de l'Assemblée nationale, porte-parole de l'opposition officielle en matière d'accès à l'information, porte-parole de l'opposition officielle en matière d'éthique, porte-parole de l'opposition officielle en matière de justice.

Les termes « principaux facteurs » pourraient aussi faire référence aux facteurs qui influencent l'efficacité d'un modèle d'IA. Ces facteurs peuvent inclure la qualité et la quantité des données utilisées pour entraîner le modèle, la complexité du modèle ainsi que les stratégies d'entraînement, d'évaluation et de maintenance du modèle. Les législateurs avait-il en tête précisément cette définition ? Dans ce cas, puisque les facteurs ne sont pas définis, chaque entreprise serait donc libre de choisir ceux qu'elle veut divulguer au public. Cela laisserait une grande marge de manœuvre aux entreprises et perdrait du sens auprès des consommateurs qui souhaiteraient comparer deux solutions d'IA pour un même service.

Les termes « principaux facteurs » sont très ambigus et portent à confusion. Des professionnels possédant une expertise dans le domaine de l'IA pourraient en faire des interprétations différentes et les adapter selon les objectifs de l'entreprise et non selon les intérêts du consommateur. Cette diversité de sens devrait encourager les législateurs à limiter l'usage de ces termes à sens multiples et à définir précisément les mots utilisés dans les textes de loi.

3.5.3 Définition du terme « paramètres » dans le contexte de l'IA

Le terme « paramètres », en langage informatique regroupe les variables internes à l'algorithme qui sont continuellement ajustées pendant le processus d'apprentissage.

Imaginons une balance que l'on souhaite utiliser pour peser des objets. Avant de commencer, nous devons nous assurer que la balance est bien équilibrée à zéro lorsqu'il n'y a aucun objet dessus. Cette étape d'initialisation permet à la balance de calibrer son mécanisme interne, associant ainsi la valeur zéro à un état. Une fois cette calibration effectuée, la balance est prête à peser des nouveaux objets. Ce processus illustre comment un ajustement des paramètres (ici, l'équilibrage à zéro) est essentiel pour que le système fonctionne correctement par la suite. Ce principe est comparable à l'ajustement de chacun des paramètres dans un modèle d'intelligence artificielle pendant sa phase d'apprentissage qui permet ensuite de prédire de nouvelles valeurs.

Lors de l'entraînement d'une IA, les paramètres sont ajustés automatiquement pour minimiser l'erreur entre les prédictions de l'algorithme et les données cibles réelles d'entraînement. Cette optimisation permet à l'algorithme d'apprendre des patrons dans les données afin d'émettre des prédictions sur de nouvelles données (par exemple, des données de tests, de simulations ou de production) similaires à celles observées durant la phase d'apprentissage. Les paramètres les plus communs aux algorithmes d'IA sont les poids. Dans un réseau de neurones, les poids des connexions déterminent l'importance et la contribution de chaque entrée pour donner une sortie. Imaginez chaque neurone comme un poste de contrôle qui reçoit des signaux ; les poids ajustent la force de ces signaux. Pendant l'apprentissage, le réseau ajuste ces poids pour mieux prédire ou comprendre les données, un peu comme apprendre la recette parfaite en ajustant les quantités d'ingrédients. Tout comme le concept d'hyperparamètres exploré à la section 5.3.2, les paramètres sont basés sur des notions mathématiques très poussées et complexes. Certains algorithmes peuvent posséder des milliards de paramètres. Par exemple, GPT-3 développé par OpenAI possède un total impressionnant de 175 milliards de paramètres (Brown et al., 2020). Expliquer la signification et la valeur de ceux-ci au consommateur moyen est impossible, en plus d'être impertinent pour sa compréhension du fonctionnement de l'algorithme. Au contraire, le Canada recommande que, lors de la publication de contenu destiné à un public général, celui-ci doit être présenté dans un format dont l'indice de lisibilité Flesch-Kincaid[2] est de 8 ou moins, ce qui correspond à un niveau de compréhension de l'enseignement de deuxième secondaire (Public Works and Government Services Canada, 1997).

3.5.4 Degré de gravité de l'impact d'un système d'intelligence artificielle

L'intuition selon laquelle un système d'IA devrait fournir des explications sur ses prédictions est particulièrement pertinente dans plusieurs contextes. Prenons l'exemple d'une procédure légale : l'utilisation d'une IA pour évaluer si un prisonnier devrait être éligible à une libération conditionnelle, en fonction de son risque de récidive. Dans ce scénario, les enjeux sont élevés.

---

[2] Outil qui permet d'approximer, en fonction de la longueur des mots et des phrases, un degré de facilité de compréhension d'un texte.

Une mauvaise prédiction peut soit mettre en danger la société en libérant un individu dangereux, soit injustement prolonger l'incarcération d'une personne ne présentant pas un risque significatif. Dans un tel contexte, il est important que les décisions prises par l'IA soient non seulement précises, mais aussi compréhensibles pour les personnes concernées, y compris les détenus, les juristes et les juges.

Ces observations sont valables pour plusieurs systèmes d'IA, particulièrement pour ceux opérant dans les domaines juridiques, de la santé et de la finance. Cependant, il est aussi possible de nommer plusieurs contextes où l'on utilise des données personnelles pour créer des systèmes d'IA, mais où une explication n'est pas nécessaire du point de vue du consommateur.

Examinons l'exemple d'un système d'IA complètement automatisé, que l'on pourrait retrouver sur Facebook, qui prédit le degré de compatibilité amoureuse. Les utilisateurs permettent au système d'IA d'utiliser certaines de leurs données personnelles présentes sur la plateforme. Par contre, les utilisateurs n'expriment aucun souhait de connaître le raisonnement menant au taux de compatibilité amoureuse. La fonctionnalité est perçue par les utilisateurs comme appréciable mais surérogatoire. Ce n'est donc pas une nécessité. Le même principe peut être appliqué aux systèmes de recommandations, typiquement aperçus sur des plateformes de diffusion telles que Netflix ou encore Spotify. Celles-ci utilisent des données personnelles pour émettre des recommandations. En revanche, les utilisateurs ne donnent pas beaucoup d'importance aux raisonnements utilisés pour fournir ces recommandations. Ils se fient plutôt à la pertinence de la recommandation.

D'une autre part, il est possible d'imaginer un système d'IA qui n'utilise pas de données personnelles, mais pour lequel des utilisateurs souhaiteraient avoir une explication de la logique de l'algorithme. Dans ce cas, l'explication ne serait pas requise par la loi. Un exemple serait un système d'IA utilisé pour la prédiction de risques géologiques. Le système d'IA analyserait des données telles que la composition du sol, les précipitations, les modèles climatiques, et les données historiques de mouvements de terrain pour prédire les zones à risque. Les résultats, basés sur des données principalement géophysiques et non personnelles,

pourraient forcer les habitants à déménager pour des risques de sécurité. Dans un tel cas de figure, il est plausible de supposer que les habitants voudraient avoir accès à la logique utilisée par l'algorithme avant d'être relocalisés.

Ce que nous relevons de ces exemples est que l'obligation de fournir des explications sur le raisonnement de l'algorithme ne devrait pas dépendre que de l'utilisation de données personnelles, mais aussi du degré d'impact de la prédiction sur la vie ou le quotidien de la personne affectée par la prédiction. Les demandes relatives à l'article 12.1 peuvent sembler extrêmes pour certains systèmes qui sont, d'une perspective d'ingénierie et pour les utilisateurs finaux, non critiques.

3.6 Considérations et recommandations pour la mise ne place pratique des lois québécoises portant sur la XIA

Cette section propose des solutions pour résoudre certains problèmes liés à la réglementation du XIA au Québec.

Dans le contexte des lois québécoises, il n'était pas envisageable d'évaluer la pertinence de nos recommandations à l'aide de méthodologies de test traditionnelles telles que le test A/B, généralement considéré comme la méthode la plus rigoureuse en matière d'évaluation empirique. Ce type de test convient toutefois principalement à des interventions ou politiques ciblées et circonscrites, tandis qu'il s'avère inapplicable pour des recommandations à plus haut niveau, comme celles formulées dans le cadre de la présente recherche.[3] En revanche, ces recommandations résultent d'une analyse interdisciplinaire de perspectives juridiques, d'ingénierie de l'IA et d'éthique des technologie récoltées grâce à plusieurs leviers.

[3] Dans un environnement de type « bac à sable », un test A/B pourrait néanmoins être envisagé lorsqu'il s'agit d'un programme expérimental ou d'un projet de recherche s'échelonnant sur plusieurs années, permettant la mise en place d'un dispositif d'évaluation longitudinal. Ce n'était toutefois pas envisageable dans le cadre temporel d'un projet de maîtrise.

1. Une revue de la littérature scientifique sur la XIA. Elle inclue notamment les travaux de Chamola et al. (2023), Saeed & Omlin (2023), Liao & Varshney (2021) et les lignes directrices éthiques de la Commission européenne (2019), qui mettent en avant des concepts comme la proportionnalité de l'explication, la lisibilité des modèles, et la distinction entre explication locale et globale.
2. Une analyse juridique de l'article 12.1 de la *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels* (LPRPSP). Celle-ci a été présentée dans les sections précédentes, identifiant des flous terminologiques et des défis de mise en œuvre pour les ingénieurs.
3. Des constats empiriques issus du Rapport du sondage sur la Loi 25 (2023). Ceux-ci révèlent une préoccupation généralisée des entreprises quant au manque de clarté de la loi, ainsi qu'un faible niveau de compréhension du fonctionnement de l'IA dans la population québécoise (seuls 13 % déclarent savoir précisément ce qu'est l'IA).
4. Des discussions avec des professionnels du domaine de l'ingénierie de l'IA ainsi que des experts juridiques. Ces discussions ont enrichi la réflexion, en mettant en lumière les difficultés pratiques rencontrées sur le terrain pour concilier performance algorithmique et obligations de transparence, ainsi qu'en proposant des solutions pour palier ses difficultés.
5. L'intégration de normes internationales. Les recommandations s'inscrivent dans la continuité des normes ISO/IEC récentes portant sur l'intelligence artificielle, particulièrement les normes ISO/IEC 22989:2022 (Artificial Intelligence – Concepts and terminology), ISO/IEC 24028:2020 (Overview of trustworthiness in AI) et ISO/IEC 23894:2023 (Guidance on risk management).

3.6.1 Recommandations à l'intention des autorités législatives et réglementaires québécoises

Dans un premier temps, il apparaît essentiel que le législateur clarifie la terminologie employée dans l'article 12.1 de la Loi sur la protection des renseignements personnels dans le secteur privé. Les termes tels que « raisons », « principaux facteurs » ou « paramètres » souffrent d'un manque de définition, ce qui crée des zones d'interprétation floues tant pour les entreprises que

pour les autorités de régulation. Une telle clarification pourrait s'appuyer sur les définitions issues de la norme ISO/IEC 22989:2022, qui établit un vocabulaire harmonisé pour l'intelligence artificielle, et devrait idéalement prendre la forme d'un glossaire technique.

Par ailleurs, il est recommandé d'intégrer explicitement la notion de « compréhensibilité » dans l'obligation d'explication imposée aux entreprises. Une explication techniquement valide mais inaccessible pour un citoyen non spécialiste ne satisfait ni l'intention de la loi ni les principes fondamentaux de transparence. Cette exigence de lisibilité est soutenue par d'autre entité de régulation et de normalisation tel que les lignes directrices européennes sur l'IA de confiance, ainsi que par la norme ISO/IEC 24028 :2020, qui souligne que les systèmes dignes de confiance doivent être compréhensibles pour leurs utilisateurs.

Une troisième recommandation centrale concerne l'introduction d'un principe de proportionnalité en matière d'explicabilité. L'article 12.1 s'applique uniformément à tous les systèmes de traitement automatisé de données personnelles, sans considération du degré d'impact des décisions rendues par l'IA. Or, la littérature éthique et les standards tels que l'ISO/IEC 23894:2023 insistent sur l'importance d'une approche graduée en fonction du niveau de risque. Ainsi, les exigences d'explication devraient être modulées en fonction de la sensibilité des décisions automatisées et de leurs conséquences potentielles sur les droits et libertés des individus.

Dans la perspective de faciliter l'interprétation de la Loi 25 et d'accompagner son application progressive, il serait également pertinent que ces éléments soient intégrés sous un format de lignes directrices officielles émises par une autorité administrative compétente. Ces lignes directrices constitueraient un outil d'interprétation du cadre légal, utilisable tant par les entreprises que par les autorités de régulation, voire par les tribunaux dans certaines circonstances. Leur caractère non contraignant leur confère une souplesse d'adaptation précieuse, particulièrement dans un domaine comme celui de l'intelligence artificielle, où les technologies, les usages et les risques évoluent rapidement. Toutefois, il importe de souligner que cette flexibilité comporte également certaines limites. Lorsqu'elle est excessive ou sujette

à des révisions fréquentes, elle peut créer une incertitude réglementaire qui désavantage les petites et moyennes entreprises, lesquelles disposent de ressources juridiques plus limitées pour assurer un suivi constant des modifications. À l'inverse, les grandes organisations, mieux dotées sur le plan légal et administratif, tirent souvent parti de cette marge d'interprétation accrue. Ainsi, la souplesse des lignes directrices ne doit pas être perçue comme une invitation à les modifier continuellement, mais plutôt comme un mécanisme d'ajustement raisonné, permettant d'assurer un équilibre entre prévisibilité, équité et adaptabilité du cadre d'application de la Loi 25.

3.6.2 Recommandations à l'intention des entreprises et développeurs de système d'IA

Dans le contexte actuel, marqué par un cadre juridique encore flou et par l'absence d'exemples jurisprudentiels ou administratifs concrets sur la façon de se conformer à l'article 12.1 de la Loi 25, les entreprises québécoises se retrouvent dans une situation d'incertitude. Sans directives opérationnelles claires ni précédents sur les bonnes pratiques à adopter, il devient difficile de faire des choix éclairés. Pourtant, ces entreprises demeurent légalement tenues de s'assurer que leurs systèmes d'IA sont conformes aux exigences d'explicabilité, de transparence et de rectification prévues par la loi. Dans ce contexte, les experts juridiques recommandent qu'elles s'appuient sur les standards internationaux reconnus, sur les pratiques de l'ingénierie responsable, ainsi que sur les principes d'éthique algorithmique largement acceptés afin de démontrer leur diligence raisonnable et leur volonté de conformité.

Les entreprises appelées à concevoir ou utiliser des systèmes d'IA devraient adopter une approche différenciée dans leur manière de fournir des explications, en tenant compte du profil de l'utilisateur et du contexte d'utilisation. Plus précisément, il serait pertinent de privilégier des explications locales (personnalisées) pour les cas où les décisions ont un impact significatif, et d'opter pour des explications globales (généralisées) lorsque l'enjeu est moindre. Cette distinction repose sur des bases scientifiques, notamment les travaux de Chamola et al. (2023), Saeed & Omlin (2023), Liao & Varshney (2021), présentant les travaux

de LIME et SHAPE, tout en s'inscrivant dans les recommandations de la norme ISO/IEC 24028 relative à l'adéquation de l'explication au public cible.

En outre, les entreprises devraient systématiser la documentation de leurs modèles, incluant les données d'entraînement, les versions d'algorithmes, les paramètres modifiés et les choix techniques relatifs à l'explicabilité. Cette pratique permettrait de répondre plus efficacement aux demandes de rectification prévues par la Loi 25, tout en renforçant l'audibilité des systèmes. Elle s'inscrit pleinement dans les exigences de traçabilité formulées dans les normes ISO/IEC 24028 :2020 et 23894 :2023.

Traditionnellement, le niveau de rigueur documentaire attendu varie en fonction de la taille de l'entreprise. Toutefois, plusieurs cadres récents en gouvernance de l'IA (par exemple, la norme ISO/IEC 23894 :2023 et les approches de gestion du risque proposées dans l'AI Act européen) suggèrent une évolution vers une graduation des exigences basée sur la criticité des systèmes plutôt que sur la structure organisationnelle. Cette perspective permet d'assurer une proportionnalité entre les efforts de traçabilité et les risques concrets associés à chaque projet d'intelligence artificielle, favorisant ainsi une conformité plus juste et efficace.

Dans les domaines critiques comme la santé, le droit ou les finances, il est également recommandé de favoriser, lorsque cela est possible, des algorithmes intrinsèquement interprétables. Bien que les modèles complexes, dits « boîtes noires », soient souvent plus performants, ils ne permettent pas toujours d'offrir des explications fiables et compréhensibles. En contexte sensible, les modèles de type arbre de décision, régression logistique ou règles experts constituent des alternatives plus responsables, capables de respecter à la fois les exigences de performance et les normes d'équité procédurale.

Enfin, dans les cas où une entreprise développe ou déploie un système d'IA dont les décisions pourraient avoir des effets significatifs sur les droits fondamentaux, la sécurité, ou la vie des personnes concernées, il est fortement recommandé de consulter un expert juridique spécialisé en protection des renseignements personnels ou en droit des technologies. Une telle

collaboration interdisciplinaire permettrait de mieux anticiper les risques de non-conformité et d'ajuster, dès la phase de conception, les mesures d'explicabilité aux exigences légales en vigueur.

3.6.3 Recommandations à l'intention des organismes publics de soutien et d'encadrement

Les institutions québécoises qui accompagnent les entreprises dans leur transformation numérique, comme Investissement Québec, le CEIMIA ou le ministère de l'Économie, de l'Innovation et de l'Énergie, devraient élaborer des guides pratiques sectoriels portant sur la mise en œuvre de l'explicabilité en IA. Ces outils pourraient prendre la forme de fiches techniques adaptées à différents contextes d'usage (ex. : ressources humaines, transport, marketing) et seraient conçus pour être accessibles aux PME, qui sont souvent démunies face aux nouvelles exigences légales. Cette approche s'inspire des initiatives européennes en matière d'évaluation d'impact algorithmique et des recommandations de l'OCDE sur la gouvernance des systèmes intelligents.

En complément, il serait souhaitable que ses organismes publics de soutien et d'encadrement ou le gouvernement du Québec mettent en place des « bacs à sable réglementaires » (regulatory sandbox) permettant aux entreprises de tester, dans un environnement supervisé, des technologies émergentes intégrant des modules d'explication. Un bac à sable réglementaire désigne un environnement supervisé dans lequel les entreprises peuvent expérimenter des systèmes d'IA innovants, incluant des modules d'explication, tout en étant accompagnées par les autorités réglementaires. Il permet ainsi de concilier innovation technologique et respect des exigences légales, en créant un espace d'expérimentation où les règles peuvent être appliquées de manière graduelle et flexible. Ce type de dispositif, encore peu connu au Québec, a démontré son efficacité dans d'autres juridictions. Par exemple, en Espagne, un premier bac à sable réglementaire en IA a été lancé en 2024 (White & Case, 2024) et en Finlande, un tel dispositif sera opérationnel dès février 2026 (Aholainen, 2025), dans le cadre de la transposition du Règlement européen sur l'IA qui impose que tous les états membres de l'Union européenne dispose d'au moins d'un bac à sable réglementaire d'ici août 2026. Ces

dispositifs permettent aux entreprises participantes de tester leurs modèles dans un cadre protégé, de collaborer directement avec les autorités, et de produire une documentation de conformité qui peut servir ultérieurement devant les tribunaux ou les organismes de contrôle.

3.6.4 Recommandations à l'intention de la communauté scientifique et académique

La communauté académique, tant en sciences informatiques qu'en droit, en communication, des sciences cognitives et de la conception numérique, est invitée à collaborer à l'élaboration d'un glossaire interdisciplinaire québécois des termes relatifs à la XIA. Des référentiels d'interprétation opérationnels qui serviraient de pont entre les prescriptions juridiques et les capacités réelles de l'ingénierie pourraient aussi être très utile. Ces initiatives permettraient de jeter les bases d'une culture commune de l'intelligence artificielle explicable, en facilitant la compréhension mutuelle entre ingénieurs, juristes, décideurs politiques et citoyens.

Une autre orientation prometteuse consisterait à créer un laboratoire dédié à l'explicabilité, où des prototypes d'IA seraient testés auprès de divers profils d'utilisateurs dans des contextes réels ou simulés. Cette approche, déjà mobilisée dans les domaines de la santé ou de la mobilité urbaine, permettrait de collecter des données qualitatives et quantitatives sur la réception des explications algorithmiques, leur lisibilité, leur influence sur la compréhension ou la confiance, et les biais cognitifs qu'elles peuvent engendrer. Ces travaux permettraient d'identifier ce que les utilisateurs considèrent comme une explication « compréhensible » au sens pratique et juridique. Le laboratoire pourrait contribuer à combler l'écart entre la conformité et la pratique.

De plus, dans la mesure où l'article 12.1 ouvre potentiellement la voie à un droit à l'explication, sans en définir clairement la portée, la communauté scientifique pourrait participer à l'élaboration de normes interprétatives souples, telles que des lignes directrices interdisciplinaires, visant à opérationnaliser ce droit. Ces travaux pourraient être réalisés en collaboration avec les régulateurs, et constituer un appui technique utile pour les tribunaux, à la manière des lignes directrices utilisées en droit administratif ou en matière de vie privée.

3.7 Conclusion : vers une harmonisation des lois québécoises avec les avancées technologies en XIA

Cette étude visait à mieux comprendre les enjeux juridiques et technologiques soulevés par l'article 12.1 de la Loi 25, dans un contexte où l'intelligence artificielle explicable (XIA) s'impose comme un levier pour instaurer la confiance envers les systèmes complètement automatisés qui utilisent des données personnelles. En croisant les perspectives du droit et de l'ingénierie, l'article a mis en lumière les limites actuelles de la législation québécoise en matière d'explicabilité algorithmique.

Nos analyses ont révélé plusieurs tensions importantes : un vocabulaire juridique imprécis susceptible de générer des interprétations divergentes, l'application uniforme d'obligations d'explication indépendamment du niveau de risque des systèmes concernés, et des attentes légales qui dépassent parfois les capacités techniques actuelles des outils de la XIA. Ces constats soulignent un risque d'insécurité juridique pour les entreprises, ainsi que des difficultés concrètes pour les concepteurs d'IA cherchant à respecter les exigences de la loi.

L'article contribue à combler ces écarts en formulant des recommandations fondées sur des standards internationaux et adaptées aux réalités du terrain. Celles-ci offrent un cadre d'action clair à la fois pour les législateurs, les entreprises, les organismes publics et la communauté scientifique, en suggérant notamment l'élaboration de lignes directrices interprétatives, souples et évolutives.

Au-delà des propositions techniques, nous appelons à renforcer la collaboration interdisciplinaire entre juristes, ingénieurs, chercheurs et régulateurs. Une telle approche collective est indispensable pour traduire les principes de transparence et de responsabilité en pratiques applicables, compréhensibles et efficaces.

CHAPITRE 4

LE DÉVELOPPEMENT D'UNE IA EXPLICABLE : ENTRE PRINCIPES ÉTHIQUES GÉNÉRAUX ET MESURES CONCRÈTES

Camélia Raymond[a], Marc-Kevin Daoust[b], Sylvie Ratté[c]

[a, b, c] Département de Génie Logiciel et des TI, École de Technologie Supérieure
Département des Enseignements Généraux, École de Technologie Supérieure
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article publié dans « Dialogue: Canadian Philosophical Review / Revue canadienne de philosophie », juillet 2025

4.1 Résumé

Les ingénieurs en IA ont besoin de directives applicables pour l'implémentation de principes éthiques dans leurs solutions technologiques. Mais comment y arriver ? Dans cet article, nous prenons le cas du développement de l'intelligence artificielle explicable (XIA) comme point de départ. Sur le plan des mesures concrètes devant être intégrées à l'IA pour la rendre explicable, nous remettons en question l'approche universaliste. Nous proposons une méthodologie normative pour évaluer les mesures de la XIA adaptées à des contextes spécifiques. Cette approche intègre l'éthique dans le développement de l'IA, offrant ainsi une méthode pragmatique pour les ingénieurs, régulateurs et chercheurs en éthique.

4.2 Abstract

AI engineers need applicable guidelines for implementing ethical principles into their technological solutions. But how can this be achieved? In this article, we take the development of Explainable Artificial Intelligence (XAI) as our starting point. First, we challenge the universalist approach, the view according to which some measures are necessary or sufficient for XAI in every context. Then, we propose a normative methodology for evaluating XAI measures that is adapted to specific contexts. This approach better integrates ethics into AI

development by offering an adaptable and pragmatic method for engineers, regulators, ethical researchers, and decision-makers.

4.3 Introduction

Les attentes sociales et commerciales envers les ingénieurs logiciels ne se limitent pas à la compétence technique. Dans le monde de l'intelligence artificielle (IA), les ingénieurs doivent aussi incorporer la dimension éthique dans leurs travaux. Historiquement focalisés sur l'innovation technique et la résolution de problèmes complexes, les ingénieurs en IA sont aujourd'hui confrontés à une responsabilité grandissante vis-à-vis des effets sociaux des technologies. Cette évolution est largement motivée par une pression sociale croissante pour un développement technologique responsable, qui considère les impacts sociétaux (Garrett et al., 2020).

Parallèlement à cette prise de conscience, un écart notable persiste entre les compétences techniques acquises par les ingénieurs et leurs connaissances en éthique. Bien que les cursus en ingénierie intègrent des principes éthiques, les exemples concrets liés au développement logiciel, et plus spécifiquement à l'IA, demeurent restreints (Garrett et al., 2020). Les développements récents dans les programmes universitaires tentent de combler cette lacune, mais une grande partie des ingénieurs actuellement en poste n'ont pas bénéficié de cette formation élargie. Cette situation soulève un problème pour les ingénieurs en IA : être tenu responsable des développements technologiques dans le domaine de l'IA, sans en comprendre complètement les enjeux éthiques. Face à ces manques en matière de formation, les ingénieurs en IA s'efforcent de trouver des solutions pratiques pour relever les défis éthiques inhérents à leur domaine

La nécessité d'augmenter la transparence des IA, c'est-à-dire d'en faciliter la compréhension et l'accès à l'information, s'inscrit dans ce tournant. Un effort collectif émerge pour développer et intégrer des techniques d'intelligence artificielle explicable (XIA). Cet engagement se reflète dans les réglementations proposées par le Canada, les États-Unis ou encore la

Commission européenne, qui établissent des exigences spécifiques concernant la XIA. Par exemple, voici comment la Commission européenne définit cette dernière :

« L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'intelligence artificielle et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'intelligence artificielle). L'explicabilité technique suppose que les décisions prises par un système d'intelligence artificielle peuvent être comprises et retracées par des êtres humains. [...] Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur) » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

Il y a deux grandes tendances dans la recherche sur la XIA. La première peut être qualifiée d'universelle, au sens où elle cherche à établir des principes et des mesures concrètes qui sont valables dans tous les contextes possibles. La seconde peut être qualifiée de contextuelle, au sens où elle part du principe que les mesures concrètes à implémenter dans une organisation vont, dans une large mesure, dépendre des caractéristiques ou faits de la situation (par exemple : les types d'utilisateurs de la plateforme et l'environnement dans lequel la solution est déployée).

Cet article poursuit deux objectifs. D'abord, en ce qui a trait à la XIA, nous remettons en question l'approche des solutions universelles. Le rejet des approches universalistes soulève toutefois un problème. Si nous souhaitons bien guider les ingénieurs voulant intégrer l'éthique à leur travail, que pouvons-nous leur dire de plus que « ce qu'il faut faire dépend du contexte » ? En d'autres termes, quels outils pouvons-nous fournir aux ingénieurs pour les aider à naviguer à travers les situations auxquelles ils sont confrontés ? Ce questionnement nous amène ensuite à proposer une méthodologie pratique de réflexion permettant d'évaluer l'efficacité d'une approche de la XIA pour soutenir les ingénieurs dans la prise de décisions éthiques.

Le plan de l'article va comme suit. Dans les sections 4.3 et 4.4, nous clarifions ce que nous voulons dire par approches « universaliste » et « contextualiste », et nous présentons la méthodologie de Clinton Castro et al. (2023) permettant de déterminer quelle approche est appropriée à certains principes éthiques. Puis, dans la section 4.5, nous évaluons trois mesures de la XIA avec cette méthodologie, en nous posant deux questions essentielles pour chaque paire principe éthique/mesure concrète : la mesure est-elle (1) suffisante et (2) nécessaire pour satisfaire au principe ? Nos observations indiquent que, bien qu'une mesure spécifique puisse être utile dans certains contextes, aucune mesure ne semble universellement nécessaire ou suffisante. Cela nous amène à proposer aux ingénieurs une approche devant composer avec l'incertitude générée par les différents contextes. La section 4.6 expose cette méthodologie, que l'on nomme la matrice normative pour l'évaluation des mesures en XIA. La section 4.7 présente l'application de cette matrice au moyen de scénarios fictifs dans le secteur de l'aéronautique.

En proposant cette approche, l'article enrichit la littérature sur la XIA en privilégiant l'analyse de mesures d'explicabilité adaptées aux besoins spécifiques des projets plutôt que la recherche d'une solution universelle. En introduisant la matrice normative pour l'évaluation des mesures en XIA, ce travail souligne l'importance d'une analyse contextuelle et offre un cadre méthodologique pour intégrer les considérations éthiques dans le développement de l'IA. Cette démarche vise à fournir aux ingénieurs en IA des outils pour une réflexion pratique sur l'application de l'explicabilité, contribuant ainsi à une meilleure compréhension et mise en œuvre des principes éthiques dans l'industrie de l'IA.

4.4 Aperçu des tendances universelles et contextuelles de la XIA

La recherche sur la XIA se divise en deux grandes tendances : les approches universelles et les approches contextuelles. Les approches universelles cherchent à développer des standards et mesures applicables à tous les contextes, en mettant l'accent, par exemple, sur la création d'un lexique commun, de méthodes uniformes, et de critères d'évaluation fixes pour les décisions des modèles d'IA. Par comparaison, les approches contextuelles adaptent l'explicabilité aux

spécificités des différents domaines d'application, reconnaissant que les besoins et les définitions de l'explicabilité varient selon le contexte ciblé.

La recherche d'approches universelles constitue un axe majeur des efforts scientifiques actuels. Ces recherches visent, par exemple, à établir un lexique standardisé pour le domaine de la XIA, à développer des méthodes et techniques applicables de manière globale et à concevoir des critères d'évaluation uniformes pour les approches de la XIA. Selon Othman Benchekroun et al. (2020), « La XIA est primordiale pour une IA de qualité industrielle ; cependant, les méthodes existantes ne répondent pas à cette nécessité (industrielle), en partie en raison d'un manque de standardisation des méthodes d'explicabilité » (Benchekroun et al., 2020, p. 1 ; nous traduisons). Les « standards » mentionnés ici font référence à des outils qui devraient être uniformément intégrés dans les pratiques des entreprises pour renforcer la XIA. Pradeep Reddy et Pavan Kumar (2023) identifient le manque de solutions XIA universelles et standardisées comme étant l'un des défis majeurs de l'IA. Leander Weber et al. (2023) ainsi que Christopher J. Anders et al. (2021) proposent une solution universelle de la XIA, alors que Phuong Quynh Le et al. (2023) ainsi que Mohamed Karim Belaid et al. (2022) exposent une méthode qui permet d'évaluer collectivement les solutions de la XIA. Le dénominateur commun de ces programmes de recherche est que les solutions proposées sont censées s'appliquer à tous les contextes imaginables.

La deuxième tendance de recherche, l'approche contextuelle, se présente comme un axe émergent de la XIA. De plus en plus de chercheurs se penchent sur cet axe, arguant que, malgré l'abondance de recherches, l'approche universelle peine à fournir des solutions efficaces. À l'inverse de l'approche universelle, les approches contextuelles cherchent à établir des lexiques spécifiques à des secteurs industriels (médecine, système bancaire, aviation) ou encore des lexiques dont la terminologie permet de définir le contexte. Par exemple, Rune Nyrup et Diana Robinson (2022) proposent une vision contextuelle de la XIA appliquée au domaine médical. Pour eux, les standards d'explicabilité varient en fonction, notamment, du public et des objectifs du système. Meike Nauta et al. (2023) remettent même en question l'utilité de la définition de la XIA, estimant que celle-ci devrait varier en fonction du contexte. Dans la

littérature, une critique récurrente formulée contre les algorithmes de la XIA stipule que, puisque ceux-ci sont souvent développés sans une définition claire de leur utilisation spécifique, ils sont inappropriés dans certains contextes. Q. Vera Liao et Kush R. Varshney (2021), Niels van Berkel et al. (2022) ainsi que Heike Felzmann et al. (2020) mettent en évidence que le contexte d'utilisation de ces algorithmes affecte le succès de leur déploiement auprès des utilisateurs. Aussi, Jianlong Zhou et al. (2021) soutiennent qu'il n'existe pas encore d'indicateur reconnu pour la qualité des méthodes d'explication dans la XIA, puisque celle-ci est un concept subjectif et que la qualité perçue d'une explication dépend de l'environnement et des parties prenantes.

La dualité entre les approches universelles et contextuelles de la XIA met en lumière la complexité de développer des explications d'IA qui soient à la fois précises et largement applicables.

4.5 Approche normative de justification d'une mesure concrète pour un principe éthique appliqué à l'IA

4.5.1 Des grands principes aux mesures concrètes

La distinction entre les approches universaliste et contextualiste concerne le choix des mesures concrètes à mettre en application dans une organisation ou un système. Pour bien comprendre ce point, il faut d'abord faire quelques rappels concernant les démarches d'analyse courantes en éthique des technologies.

Dans la recherche appliquée en éthique des technologies, les avis comprennent généralement au moins trois éléments : (i) une description des faits de la situation, (ii) une description des valeurs en jeu et (iii) des mesures concrètes à implémenter dans une organisation ou un système. Prenons, par exemple, les 19 avis publiés par la Commission de l'éthique en science et en technologie (CEST) de 2003 à 2024. Tous ces avis ont sensiblement la même structure. On y décrit d'abord le fonctionnement d'une technologie, ainsi que les actions déjà posées au Québec et au Canada pour encadrer la technologie en question. Puis, on esquisse différentes

valeurs pertinentes dans la situation. Finalement, on propose des mesures concrètes aux décideurs pour bien encadrer la technologie à l'étude (voir, par exemple, CEST, 2021 ; 2023a ; 2023b ; 2023c ; 2024).

Du point de vue de la pratique des ingénieurs, s'en tenir à une description de grands principes devant régir les technologies n'est pas éclairant. Il faut relier ces principes à des mesures concrètes. Les valeurs devant guider le développement de l'IA sont bien documentées dans la littérature. Plusieurs contributions récentes affirment que l'IA devrait encourager l'exercice de notre autonomie et de notre agentivité (Rubel et al., 2021), favoriser notre compréhension de ses décisions (Fleisher, 2022), tendre à développer les vertus propres à nos fonctions (van Wynsberghe, 2016), respecter notre vie privée (Nissenbaum, 2009), ne pas accentuer les dynamiques d'oppression présentes dans la société (Noble, 2018), et ainsi de suite. Le fait d'énoncer et de souligner des principes devant guider le développement des technologies est, bien entendu, un travail essentiel. Mais pour les concepteurs, la réflexion ne peut pas s'arrêter là. Ces professionnels doivent aussi articuler les relations entre, d'une part, les principes éthiques proposés et les pratiques organisationnelles ou les choix technologiques, d'autre part [1].

Il y a différentes manières de concevoir la relation entre les valeurs que l'on souhaite mettre de l'avant dans les systèmes d'IA et les mesures concrètes pour y parvenir. Selon la définition de l'approche universaliste, certaines mesures concrètes sont soit nécessaires soit suffisantes dans tous les projets d'IA pour pleinement respecter une valeur précise en jeu. Peu importe les faits de la situation (par exemple : les types d'utilisateurs de la plateforme et l'environnement dans lequel la solution est déployée), certaines mesures concrètes doivent être implémentées pour rendre compte de cette valeur. Les approches contextualistes soutiennent plutôt que

[1] L'absence de lien entre des valeurs et des mesures concrètes peut mener, entre autres, à du fairwashing. Cela s'observe lorsque des organisations affirment qu'elles implémentent des valeurs dans les technologies développées, alors que les mesures concrètes implémentées par l'entreprise n'ont aucun lien avec celles-ci. Voir notamment Aïvodji et al. (2019) sur ce point.

différentes mesures concrètes peuvent rendre compte de nos valeurs, et que les actions pertinentes à poser dépendent, dans une large mesure, des faits de la situation.

Pour rendre ces deux thèses plus concrètes, prenons l'exemple de l'avis intitulé *La gestion algorithmique de la main-d'oeuvre : analyse des enjeux éthiques*, publié en 2023 par la CEST. Dans cet avis, la CEST propose notamment d'obliger les employeurs à divulguer le recours à la surveillance électronique au travail, afin de répondre à des valeurs éthiques comme la transparence. Deux interprétations peuvent être envisagées pour comprendre la relation entre cette valeur et la mesure recommandée.

Dans une perspective universaliste, la mesure concrète proposée, c'est-à-dire l'obligation de divulguer, pourrait être perçue comme nécessaire (il n'y a pas de transparence sans cette divulgation) ou suffisante (cette divulgation garantit une transparence minimale). Autrement dit, la mesure est une condition (nécessaire ou suffisante) pour atteindre la transparence, indépendamment du contexte. Dans une approche contextualiste, cette recommandation reflète une réponse adaptée aux conditions particulières du Québec en 2023, marquées par des technologies de surveillance électronique et des attentes sociales ou juridiques propres à cette époque. Si le contexte technologique ou sociétal venait à évoluer, par exemple avec l'émergence de nouvelles technologies moins intrusives ou une transformation des attentes éthiques, cette mesure pourrait perdre sa pertinence ou être remplacée par d'autres solutions mieux adaptées. Ces deux interprétations des conclusions de la CEST sont cohérentes avec la démarche de l'organisation, qui vise à articuler des principes éthiques et des recommandations pratiques en fonction des besoins identifiés.

Chercher des mesures concrètes universelles est tentant. Après tout, ces solutions ont le mérite de « durer dans le temps », et de s'appliquer à une foule de situations. Elles permettent d'établir une base commune de mesures nécessaires ou suffisantes pouvant être importées dans toutes les organisations ou tous les systèmes. C'est pourquoi de nombreux programmes de recherche ont pour objectif d'identifier de telles mesures.

Or, est-il réaliste d'espérer trouver des mesures concrètes universelles ? Cette question a récemment fait l'objet de réflexions. Castro et al. (2023), par exemple, proposent une méthodologie de justification d'une mesure concrète, en réponse à un principe éthique dans le domaine de l'IA, soit l'égalité formelle des chances. Pour ce faire, la méthodologie de Castro et al. articule les relations entre quatre éléments, soit :

- Les faits moraux fondamentaux : Ce sont les vérités morales de base qui sous-tendent les jugements éthiques dans une théorie donnée. Par exemple, dans le domaine de l'IA, un fait moral fondamental pourrait être le respect de la vie privée des utilisateurs. Ce principe éthique guide la conception et le déploiement des algorithmes.
- Les faits empiriques pertinents : Ce sont les faits qui ont une influence sur l'évaluation morale dans une situation donnée. En IA, cela pourrait inclure la reconnaissance des biais dans les ensembles de données. Ces biais peuvent entraîner des discriminations involontaires dans les décisions prises par les systèmes d'IA, affectant ainsi leur évaluation éthique.
- Les principes intermédiaires : Ce sont des règles ou des lignes directrices qui permettent d'appliquer des faits moraux fondamentaux à des cas pratiques. Par exemple, un principe intermédiaire en IA pourrait être l'équité et la non-discrimination dans les décisions automatisées
- Les mesures concrètes : Ce sont des actions ou des pratiques précises recommandées par une théorie éthique dans une situation réelle. Dans le contexte de l'apprentissage automatique, ces mesures pourraient être la présentation d'un formulaire de consentement d'utilisation des données lors du téléchargement d'une application qui utilise l'IA.

La figure 4.1 décrit les relations unissant ces quatre notions dans la méthodologie de Castro et al.

Appliquée au principe intermédiaire de l'équité, cette méthodologie amène les auteurs à conclure que les mesures concrètes proposées par les organisations pour atteindre l'égalité des chances dans des algorithmes ne sont ni absolument incontournables ni absolument erronées. Les auteurs concentrent leur attention sur trois mesures concrètes courantes pour améliorer

l'équité dans les algorithmes, soit : (i) le traitement anonyme des dossiers des personnes, (ii) l'égalisation des chances entre tous les groupes socioéconomiques et (iii) l'analyse contrefactuelle des informations personnelles. Toutes ces mesures ont été proposées par des institutions privées ou publiques pour respecter le principe d'égalité des chances par des algorithmes. Or, selon les auteurs, aucune de ces mesures n'est absolument nécessaire ou suffisante pour respecter l'équité. L'adéquation des mesures d'égalité des chances dépend du contexte (c'est-à-dire des faits empiriques pertinents) propre à chaque situation. En d'autres termes, lorsqu'appliquée à la notion d'équité, la méthodologie proposée par Castro et al. tend à soutenir une conception contextualiste des mesures concrètes pertinentes[2].

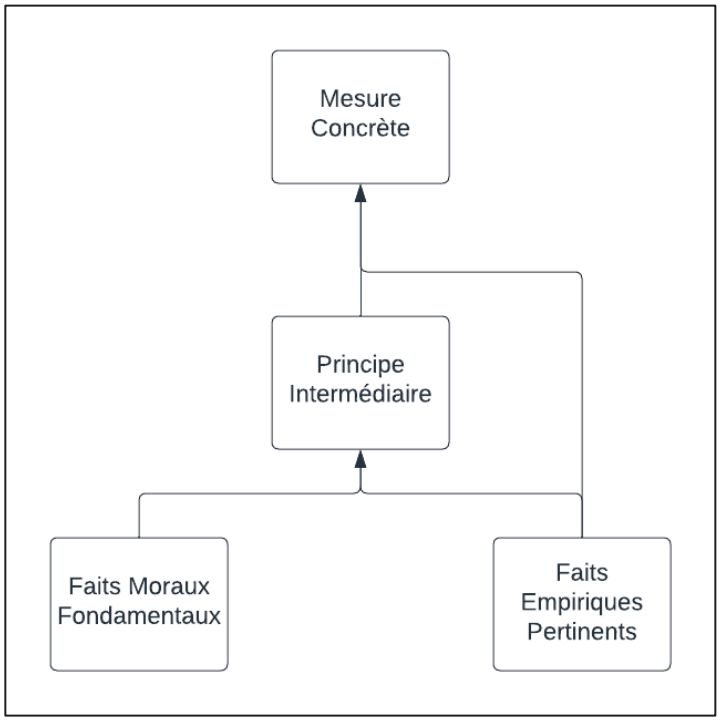


Figure 4.1 Interactions entre les concepts normatifs

[2] Les auteurs parlent de « pluralisme » des mesures concrètes.

4.5.2 La méthodologie de Castro et al. Appliquée à la XIA

La méthodologie de Castro et al. peut être appliquée aux relations entre principes intermédiaires et mesures concrètes en XIA. En suivant cette même méthodologie, nous proposons de poser la transparence comme un fait moral fondamental. La transparence, dans ce contexte, renvoie à la nécessité d'être ouvert et honnête en ce qui concerne les processus, les décisions et les actions. Elle peut être qualifiée de fondamentale, car elle touche à des valeurs morales essentielles comme la confiance, la justice et l'intégrité, des valeurs que Castro et al. identifient comme des points d'ancrage éthiques. Cette approche nous amène à définir la XIA comme principe intermédiaire, c'est-à-dire comme articulation pratique de ce fait moral fondamental dans le domaine technologique. La figure 4.2 présente notre interprétation de la méthodologie de Castro et al. appliquée à la XIA. Il restera ensuite à savoir si certaines mesures concrètes découlent universellement de ces faits moraux et principes, ou si elles doivent être ajustées en fonction des contextes donnés.

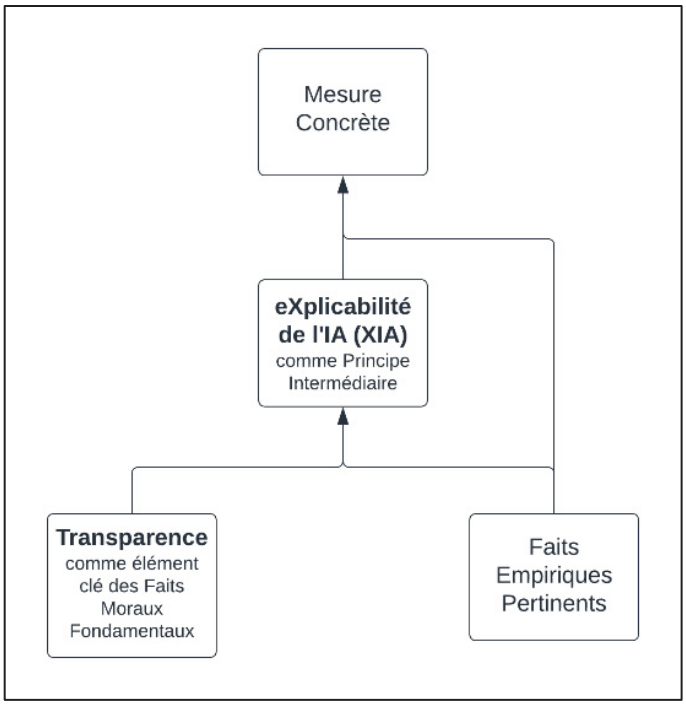


Figure 4.2 Interprétation de l'application des concepts normatifs

4.6 Évaluation normative des mesures concrètes pour le principe intermédiaire de la XIA

Dans cette section, nous souhaitons évaluer le statut (universel ou contextuel) de mesures concrètes de la XIA. Ainsi que le recommandent Castro et al. (2023), cette évaluation sera réalisée en fonction de la conformité de ces mesures avec les faits empiriques pertinents, dans un contexte donné.

Nous mettons en évidence la complexité et la diversité des approches existantes, et démontrons qu'une analyse basée sur les principes intermédiaires et les faits empiriques pertinents aide à choisir la méthode la plus appropriée pour une situation donnée. Cette évaluation pose les deux questions suivantes pour chaque paire de principes intermédiaires et de mesures concrètes :

1) Cette mesure concrète est-elle suffisante pour satisfaire au principe intermédiaire de la XIA en prenant en considération les faits empiriques pertinents étudiés ?
2) Cette mesure concrète est-elle nécessaire pour satisfaire au principe intermédiaire de la XIA en prenant en considération les faits empiriques pertinents étudiés ?

Pour répondre à ces questions, nous évaluons trois mesures concrètes, soit : (i) la publication de notes de transparence IA, (ii) la présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur, et (iii) l'affichage d'un score de similitude entre des données d'entrée. Toutes ces mesures ont été proposées par de grandes institutions ou entreprises pour respecter le principe de la XIA.

4.6.1 La publication de notes de transparence IA

Nous commençons en examinant une mesure proposée par Microsoft : la note de transparence IA. Chez Microsoft, cette note fait partie d'une initiative plus large qui vise à rendre l'IA compréhensible. Elle fournit des informations textuelles détaillées sur les services IA offerts : le fonctionnement des différents modèles d'IA au sein des services Azure de Microsoft, les choix effectués par les concepteurs de systèmes qui influencent la performance et le comportement des systèmes, ainsi que la façon dont l'ensemble des services IA de Microsoft

vont s'intégrer dans le système client, qui inclut la technologie, les personnes et l'environnement (Liao et al., 2023). Pour Azure AI Language, la note de transparence décrit les utilisations possibles des différentes fonctionnalités, telles que la reconnaissance d'entités nommées, la détection de langues, l'analyse de sentiments, et bien d'autres (Aahill et al., 2023). Chaque fonctionnalité est accompagnée de cas d'usage, de considérations sur les limites potentielles et de conseils pour améliorer la performance des systèmes. Cela inclut également des informations sur la gestion de la confidentialité des données et la personnalisation de la classification des textes.

Si l'on prenait le point de vue des autorités de régulation, il semblerait logique de s'assurer que les entreprises se conforment bien à la publication d'une note de transparence pour augmenter la compréhension d'une IA. Sur la base de cette observation, la présentation d'une note de transparence paraît nécessaire pour répondre aux attentes de la XIA.

En revanche, notons que Microsoft n'affiche pas une note de transparence pour tous les systèmes d'IA mis à la disposition de ses clients. Par exemple, l'entreprise présente une fonctionnalité de vérification des pourriels qui fait appel à l'IA pour analyser le contenu de chaque message de marketing par courrier électronique (Ferguson et al., 2024). L'IA génère une probabilité que le courrier soit un pourriel. Il n'y a pas de note de transparence pour cette fonctionnalité. Les prédictions émises par l'IA ont un faible impact sur le quotidien des utilisateurs. Les utilisateurs comprennent généralement que le système vise à améliorer l'efficacité, sans pour autant affecter significativement leur travail. Aussi, dans ce scénario, l'application de l'IA est relativement simple et bien comprise. Dans ce cas, la note de transparence est considérée comme non nécessaire. Cela indique que la note de transparence est nécessaire pour certains produits, mais pas pour tous les produits.

Lorsqu'on se penche sur la question de la suffisance, il est assez facile de voir que l'utilisation unique de la note de transparence pour tous les scénarios est loin d'être suffisante. Ce n'est pas parce que l'on connaît les principes généraux d'une IA que l'on comprend comment ceux-ci se manifestent dans un cas particulier, tel que le refus d'un prêt hypothécaire à un client.

4.6.2 La présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur

Examinons maintenant une autre mesure proposée en lien avec la XIA, soit la présentation de l'impact des caractéristiques d'entrée de l'algorithme sur la prédiction. L'application mobile Intact Assurance permet aux utilisateurs d'obtenir des rabais sur leur assurance automobile basés sur une analyse intelligente de leurs habitudes de conduite (Intact Assurance, s. d.). Cette application leur permet aussi d'observer ces habitudes et l'impact qu'elles ont sur le rabais qui leur est offert, un exemple concret de la présentation de l'impact des caractéristiques d'entrée de l'algorithme sur la prédiction.

Tout comme la mesure précédente, s'il nous était demandé, en tant qu'autorité de réglementation, de veiller à ce que les entreprises respectent effectivement l'idéal de la XIA, il semblerait raisonnable de vérifier que les systèmes fournissent des explications sur l'impact des caractéristiques d'entrée d'une IA. Pour des produits comme une police d'assurance, où les décisions de l'IA peuvent avoir un impact monétaire direct sur les utilisateurs, la compréhension de ces impacts devient cruciale pour obtenir la meilleure prime possible. Cette description de l'importance des caractéristiques serait donc jugée nécessaire pour répondre aux exigences de la XIA, ne serait-ce que pour certains produits comme une police d'assurance.

Or, cette mesure concrète est-elle nécessaire ou suffisante pour tous les produits ? Certains faits de la situation pourraient affecter notre évaluation de la satisfaction de cette mesure. Pour illustrer nos propos, prenons l'exemple du diagnostic du cancer par imagerie médicale, un système que l'on dit couramment de haute incidence.

Dans ce contexte, la nécessité de cette mesure dépend de l'objectif et de l'utilisateur visé. Pour un expert en IA ou un professionnel de la santé, la présentation de l'impact de chaque pixel de l'image pourrait être jugée nécessaire pour vérifier le fonctionnement correct du système et valider son raisonnement. En revanche, pour un patient, ces informations détaillées sont non

nécessaires, car elles ne comblent pas directement son besoin : comprendre pourquoi le diagnostic a été posé et ce que ça implique pour son traitement ou sa santé. La complexité des données, comme l'analyse de milliers de pixels, rend l'information inutilisable pour un non-expert. Le patient aurait plutôt besoin d'explications simplifiées, comme une description textuelle indiquant les principales caractéristiques qui ont influencé la décision (par exemple, la taille de la tumeur ou son emplacement), ou une visualisation intuitive mettant en évidence les zones concernées sur l'image. Ainsi, la présentation de l'impact des caractéristiques d'entrée (les pixels) sur la prédiction n'est pas une mesure nécessaire pour satisfaire les attentes de la XIA dans tous les cas d'utilisation.

Lorsque l'on se penche sur la question de la suffisance, la réponse peut être nuancée. Pour un médecin, bien que cette mesure puisse fournir des informations utiles pour analyser les décisions de l'IA, elle n'est pas suffisante à elle seule. Les médecins ont souvent besoin d'une synthèse des facteurs influents qui relie directement les données techniques aux implications cliniques, tels que la taille de la tumeur ou ses caractéristiques biologiques. Ainsi, si cette mesure peut suffire dans un cadre purement technique (par exemple, pour un expert en IA), elle doit souvent être accompagnée d'une interprétation clinique pour être pleinement exploitable dans un contexte médical.

4.6.3 L'affichage d'un score de similitude entre des données d'entrée

Finalement, examinons une dernière mesure concrète : l'affichage d'un score de similitude entre des données d'entrée. Pour comprendre de quoi il s'agit, pensons, par exemple, à un système de recommandation de films, tel que celui employé par Netflix, qui présente un score (pourcentage) de similitude entre deux films (Netflix, s. d.).

Dans ce cas, l'utilisateur interagit avec le système en choisissant des films, et le système utilise ces informations pour recommander d'autres titres similaires. Les utilisateurs bénéficient de ces recommandations personnalisées sans avoir besoin de comprendre les détails techniques sur la façon dont les recommandations sont générées. Ils sont généralement plus intéressés par

la pertinence des recommandations que par les mécanismes sous-jacents du système de recommandation. Dans ce cas, nous pourrions dire que, dans le contexte précis du choix d'un film le samedi soir, l'affichage d'un score de similitude est une mesure suffisante en ce sens qu'elle excède ce qui est requis pour permettre aux abonnés de comprendre le fonctionnement de la plateforme. Or, cette mesure n'est pas nécessaire. D'autres moyens que le score de similitude pourraient favoriser cette compréhension. Par exemple, la plateforme pourrait expliquer ses recommandations au moyen des préférences passées des utilisateurs : « Nous vous recommandons le film X car vous avez aimé le film Y. » Une telle justification permet aux utilisateurs de comprendre les recommandations, sans qu'aucun score soit mobilisé.

Examinons l'application de cette mesure concrète à d'autres contextes. Imaginons, par exemple, qu'un médecin fasse un diagnostic en présentant seulement le score de similitude entre deux patients générés par une IA. En d'autres termes, le médecin se contente de dire que le patient X a sans doute la maladie M, puisqu'il a des caractéristiques semblables à un autre patient Y aux prises avec la maladie M. Cette explication hautement limitée ne respecte pas le critère de la XIA dans le secteur médical. Dans ce cas, la mesure est nécessaire en ceci qu'elle fournit une base pour établir un diagnostic et soutient la démarche explicative du médecin. Cependant, cette mesure est insuffisante pour répondre pleinement aux attentes de la XIA, car elle ne permet pas de comprendre les raisons sous-jacentes de cette similitude. Elle ne répond pas aux questions essentielles, telles que : quelles caractéristiques précises du patient X ont contribué à cette prédiction ? Pourquoi ces caractéristiques sont-elles particulièrement importantes ?

4.6.4 Conclusion partielle

Faisons le point. Nous avons étudié trois mesures concrètes qui ont été proposées en lien avec la XIA, soit : (i) la publication de notes de transparence IA, (ii) la présentation de l'impact des caractéristiques sur la prédiction à l'utilisateur, et (iii) l'affichage d'un score de similitude entre des données d'entrée. Et nous avons conclu que celles-ci n'étaient pas absolument nécessaires ni suffisantes pour satisfaire au principe intermédiaire de la XIA dans tous les contextes.

Ces résultats mettent en évidence le caractère non universel des mesures étudiées. Leur pertinence dépend des spécificités contextuelles, telles que le type d'utilisateur, le domaine d'application ou les objectifs visés par la technologie proposée. En conséquence, les mesures concrètes possèdent un statut contextuel qui appelle une adaptation au cas par cas pour répondre aux exigences de la XIA.

4.7 Proposition d'une matrice normative pour l'évaluation des mesures en XIA

En suivant l'analyse réalisée à la section 4.6, l'un des principaux défis dans l'évaluation normative de la pertinence d'une mesure concrète pour la XIA réside dans l'identification correcte des faits empiriques pertinents. Cette analyse justifie les recherches en XIA se spécialisant dans un contexte précis.

Mais ce constat soulève un problème. Conformément à une approche contextualiste, devons-nous simplement nous dire que « tout dépend de la situation donnée » ? Pour des ingénieurs en IA qui souhaitent satisfaire à des principes éthiques intermédiaires dans leurs travaux, cette conclusion n'est d'aucun secours. Elle compromet aussi les collaborations entre des entreprises de différents secteurs désirant partager leurs pratiques. Une entreprise a beau avoir développé des pratiques satisfaisantes pour atteindre l'explicabilité dans son domaine, cela ne signifie nullement que d'autres entreprises devraient (ou même pourraient) adopter les mêmes pratiques. Ainsi, le défi consiste à rendre compte du caractère contextuel des mesures concrètes pertinentes, tout en offrant de l'accompagnement aux ingénieurs souhaitant intégrer l'éthique à leur pratique professionnelle.

Pour surmonter les difficultés que les ingénieurs rencontrent dans l'évaluation éthique de la pertinence des mesures concrètes, il est essentiel de les orienter vers une méthode permettant de préciser, relativement aux faits de la situation, lesquelles sont pertinentes. Une telle approche favorise la compréhension des enjeux et encourage une prise de décision éclairée et adaptée au contexte précis de chaque projet d'IA.

Pour progresser dans la compréhension et l'application de mesures concrètes qui soutiennent le principe intermédiaire de la XIA, nous proposons la matrice d'évaluation normative des mesures concrètes de la XIA. Cette matrice est qualifiée d'expérimentale puisqu'elle a été validée à partir de cas d'utilisation fictifs. Cette approche consiste en la création d'une matrice, qui prend en considération deux faits empiriques pertinents et leurs variations possibles. Chaque axe de la matrice représente un fait empirique différent, tandis que les cellules formées à l'intersection de ces axes reflètent les combinaisons possibles de variations de ces faits. En fonction de la position exacte d'un système dans cette matrice, une entreprise peut déterminer de manière proactive les mesures concrètes les mieux adaptées à son contexte. Le tableau 4.1 présente le gabarit d'une telle matrice.

Tableau 4.1 Gabarit de la matrice d'évaluation normative des mesures concrètes de la XIA

		Fait empirique pertinent 2		
		Variation 1	Variation 2	Variation 3
Fait empirique pertinent 1	Variation 1	Mesure concrète 1.1	Mesure concrète 1.2	Mesure concrète 1.3
	Variation 2	Mesure concrète 2.1	Mesure concrète 2.2	Mesure concrète 2.3
	Variation 3	Mesure concrète 3.1	Mesure concrète 3.2	Mesure concrète 3.3

En intégrant explicitement les faits empiriques pertinents dans la matrice, cette approche offre un outil flexible qui s'adapte aux spécificités de chaque contexte d'application. Par exemple, dans un système d'IA déployé dans le secteur médical, les solutions pourront être personnalisées en fonction de facteurs tels que la complexité des données cliniques, les besoins précis des utilisateurs (médecins, patients), ou encore l'emplacement physique de déploiement du système (urgence, centre de radiologie).

De plus, ce cadre offre un moyen structuré d'explorer les interactions entre les faits empiriques pertinents. En identifiant les combinaisons de ces facteurs dans la matrice, il devient possible d'évaluer comment les variations d'un facteur modifient l'efficacité ou la pertinence d'une mesure concrète. Cela encourage une réflexion normative proactive, car les utilisateurs de la

matrice peuvent anticiper des tensions ou des compromis éthiques liés à ces interactions, et ainsi adapter leurs choix en conséquence.

Ce cadre permet non seulement une personnalisation des solutions de XIA, mais encourage également une réflexion normative sur la manière dont divers facteurs empiriques interagissent et influencent ces solutions.

4.8 Étude de cas dans le secteur aéronautique

Pour illustrer cette approche, considérons le cas d'un constructeur d'avions. Un avion intègre de multiples systèmes aux fonctions variées, lesquels interagissent de différentes manières avec plusieurs utilisateurs. Afin d'élaborer une matrice de mesures concrètes pour la XIA, le constructeur choisit de se concentrer sur deux faits empiriques pertinents. Le premier est l'expertise de l'utilisateur du système dans le domaine aéronautique. L'utilisateur est défini en fonction du système analysé (par exemple : mécanicien, passager ou pilote) et un niveau de connaissance en IA lui est attribué avec trois variations possibles, soit basique, moyen ou élevé. Le deuxième fait empirique pertinent de cette matrice est le niveau de risque du système pour les passagers d'un avion, qui présente aussi trois variations possibles, soit minimal, limité ou élevé. Ainsi, à la suite d'une évaluation normative des faits empiriques pertinents et des mesures concrètes réalistes et accessibles, le constructeur serait amené à créer une matrice semblable à celle présentée au tableau 4.2. Pour ce faire, le constructeur aura défini, pour chaque combinaison de variations possible, une ou plusieurs mesures concrètes. Pour valider l'efficacité de cette matrice, évaluons si les mesures concrètes qui y sont présentées sont nécessaires et suffisantes pour deux systèmes d'IA que l'on pourrait retrouver à bord d'un avion. Notons que la matrice suggérée au tableau 4.2 ne sert pas de référence pour les constructeurs d'avions, mais plutôt d'exemple pour guider le format et l'applicabilité d'un tel outil.

Tableau 4.2 Matrice de mesures concrètes de la XIA en fonction de faits empiriques pertinents

		Fait empirique pertinent 2		
		Variation 1	Variation 2	Variation 3
Fait empirique pertinent 1	Variation 1	Démontrer le fonctionnement correct du système au moyen de métriques de performance universellement connues.	Démontrer le fonctionnement correct du système au moyen de métriques de performance connues du secteur d'activité global de l'aéronautique.	Démontrer le fonctionnement correct du système au moyen de métriques de performance basées sur des informations techniques approfondies sur le secteur d'activité donné.
	Variation 2	Présenter les grandes catégories de données d'entrée du système en fournissant un lexique qui explique l'utilité de chacune de ses catégories.	Présenter partiellement les données d'entrée du système. Celles qui sont généralement connues dans le domaine de l'aéronautique.	Présenter les données d'entrée qui ont mené aux prédictions, incluant des informations techniques approfondies sur le secteur d'activité donné.
	Variation 3	Fournir des scénarios « et si » de complexité moindre, expliquant ce qui pourrait se passer si différentes actions sont entreprises. Ajouter une description textuelle et visuelle.	Fournir des scénarios « et si » en temps réel des cas les plus fréquents, expliquant ce qui pourrait se passer si différentes actions sont entreprises.	Offrir à l'utilisateur l'option d'élaborer des scénarios « et si », expliquant ce qui pourrait se passer si différentes actions sont entreprises.

Le premier cas de figure que nous souhaitons explorer est le développement d'un système de gestion optimisée de la consommation de carburant. Ce système utilise des algorithmes d'apprentissage automatique pour optimiser la consommation de carburant de l'avion en calculant les trajectoires de vol les plus efficaces en temps réel, prenant en compte les conditions météorologiques actuelles, le poids de l'avion, et d'autres variables.

Ce cas de figure peut réduire les coûts d'exploitation pour les compagnies aériennes et diminuer l'empreinte carbone des vols. Ce système n'affecte pas directement la sécurité des passagers, mais améliore l'efficacité et la durabilité des opérations aériennes. Les passagers bénéficient indirectement de ces améliorations par des coûts potentiellement réduits et une conscience accrue de l'empreinte environnementale de leur voyage. Les ingénieurs chargés du développement de ce système pourraient déterminer que celui-ci présente un niveau de risque minimal pour les passagers et que l'expertise de ces derniers en matière aéronautique est limitée. En suivant la matrice des mesures concrètes de la XIA en fonction des faits empiriques pertinents (tableau 4.2), qui aurait préalablement été construite par la compagnie d'aviation, une mesure concrète de la XIA consisterait à démontrer le fonctionnement correct du système au moyen de métriques de performance universellement connues. Cela pourrait se traduire, par exemple, par l'affichage en temps réel, sur l'écran des passagers, des économies de carburant réalisées grâce au système durant le vol.

Bien que l'expertise en aéronautique des passagers soit limitée, fournir des informations compréhensibles et directement liées à l'efficacité du système favorise l'appréciation de la technologie et de ses avantages environnementaux. Cela peut également contribuer à une prise de conscience plus large des efforts de durabilité dans le secteur aérien. En ce sens, la mesure est nécessaire pour renforcer la confiance et la compréhension publiques quant à l'utilisation des technologies d'IA dans l'amélioration des opérations aériennes. De plus, l'intérêt principal des passagers envers ce système concerne l'efficacité et les implications environnementales des vols qu'ils font. Ainsi, la présentation des économies de carburant réalisées grâce au système peut être considérée comme une mesure suffisante de la XIA. Cette information répond directement à la préoccupation des passagers concernant la durabilité et l'impact

écologique de leurs voyages, en fournissant une mesure tangible des efforts de la compagnie aérienne pour minimiser son empreinte carbone.

Un deuxième cas de figure illustrant notre approche serait le développement d'un système de maintenance prédictive pour les avions. Ce système d'IA analyse les données issues des capteurs installés sur diverses composantes de l'avion, telles que les moteurs, les systèmes hydrauliques et les équipements électriques, pour prédire les pannes potentielles avant qu'elles ne se produisent. En s'appuyant sur l'apprentissage automatique et l'analyse prédictive, le système identifie les signes avant-coureurs de défaillance, permettant aux équipes d'entretien d'intervenir de manière proactive, plutôt que réactive. Cela aide à éviter les retards et les annulations de vols dus à des problèmes techniques inattendus, améliorant ainsi l'expérience globale des passagers et la fiabilité des opérations aériennes. Bien que ce système ne soit pas directement lié à la sécurité des vols en temps réel, il joue un rôle crucial dans la prévention des incidents et assure le bon fonctionnement de l'avion. La maintenance prédictive contribue à une meilleure gestion des ressources et à une réduction des coûts pour les compagnies aériennes, tout en augmentant la satisfaction des passagers grâce à une diminution des perturbations de voyage. Les ingénieurs chargés du développement de ce système pourraient déterminer que celui-ci présente un niveau de risque limité pour les passagers et que l'expertise des utilisateurs du système (techniciens et ingénieurs mécaniques) en aéronautique est élevée. En se référant à la matrice définie au tableau 4.2, une mesure concrète de la XIA pourrait être de présenter les données d'entrée qui ont mené à ces prédictions, incluant des informations techniques approfondies, adaptées au niveau d'expertise élevé des utilisateurs. Ceci pourrait être réalisé au moyen de rapports fournis aux utilisateurs du système.

La présentation des données d'entrée qui ont conduit aux prédictions du système, incluant des informations techniques approfondies, est indéniablement nécessaire pour les mécaniciens. Cette nécessité découle du rôle clé de ces derniers dans la prévention des défaillances et dans le maintien de la sécurité et de la fiabilité des appareils. L'accès à des informations détaillées leur permet de comprendre les fondements des alertes de la maintenance prédictive, facilitant ainsi l'identification rapide et précise des problèmes potentiels et la planification des

interventions. Pour les mécaniciens, la présentation des données d'entrée ayant mené aux prédictions du système est potentiellement suffisante dans leur contexte. Ceux-ci sont des experts dans le domaine aéronautique et sont familiarisés avec les nuances techniques complexes des systèmes aériens. Donc, l'évaluation de la suffisance prend en compte leur haut niveau de compétence et leur capacité à interpréter des données complexes pour diagnostiquer et prévenir efficacement les problèmes techniques.

Par le biais de ces exemples, nous pouvons valider la pertinence de la matrice des mesures concrètes de la XIA en fonction des faits empiriques pertinents. Elle permet aux ingénieurs d'avoir des lignes directrices établies et approuvées par l'entreprise, élaborées en tenant compte des faits empiriques jugés pertinents pour le secteur d'activité concerné. Elle permet aussi aux ingénieurs de différentes organisations d'évaluer la pertinence d'adopter des mesures communes et de partager leur expertise. Il leur suffit de déterminer si les faits pertinents de leurs organisations respectives sont suffisamment similaires.

Il convient d'établir que plusieurs améliorations doivent être envisagées avant d'industrialiser cette approche. Par exemple, plusieurs autres faits pertinents pourraient être pris en considération, tels que l'expertise en IA des utilisateurs ou encore le temps de compréhension accessible à l'utilisateur du système. Dans le cas d'un système d'assistance de pilotage pour des situations d'urgence, le pilote n'aurait que quelques secondes pour comprendre les explications qui lui sont fournies, alors que dans le cas du système de maintenance prédictive pour les avions, le mécanicien dispose de plus de temps. Des matrices supplémentaires ou à dimensions plus grandes devraient être explorées. En outre, la réussite de l'approche dépend fortement de la volonté et de la capacité des organisations d'investir dans des évaluations normatives et d'adapter leurs pratiques. Dans certains cas, la pression pour une mise sur le marché rapide pourrait les pousser à négliger ces considérations éthiques au profit de l'efficacité opérationnelle.

4.9 Conclusion

Dans cet article, nous avons exploré les dimensions normatives de la XIA, avec un accent particulier sur les défis que les ingénieurs IA rencontrent dans l'intégration de mesures concrètes. Pour ce faire, notre étude a abordé les deux questions suivantes :

1) Les mesures de la XIA peuvent-elles être standardisées universellement, ou doivent-elles être adaptées spécifiquement à chaque contexte d'application ?
2) Quelles méthodes et quels outils peuvent aider les ingénieurs à intégrer efficacement des considérations éthiques de la XIA dans le développement de l'IA ?

Notre recherche a d'abord mis en évidence que le choix d'une mesure de la XIA plutôt qu'une autre doit être adapté à chaque contexte pour être véritablement efficace. Cette conclusion est soutenue par l'analyse de différentes applications de la XIA dans des contextes industriels variés, où les exigences et les impacts de l'IA diffèrent grandement. Bien que des standards universels puissent fournir un cadre général, une personnalisation de ces mesures est nécessaire pour répondre précisément aux besoins éthiques, techniques et opérationnels de chaque projet. En soi, documenter la pertinence de certaines pratiques dans un contexte donné représente une avancée dans la discussion collective sur la XIA.

Pour aider les ingénieurs à intégrer efficacement l'explicabilité dans le développement de l'IA, notre recherche a proposé et détaillé la mise en œuvre d'une matrice d'évaluation normative des mesures concrètes de la XIA. Cette matrice, qui devrait être produite en amont du développement des systèmes d'IA, guide les ingénieurs à travers un processus structuré pour définir les mesures concrètes en réponse aux principes éthiques dans des situations variées. En fournissant un cadre méthodologique, la matrice permet d'augmenter la conformité aux normes éthiques et d'adapter les mesures concrètes de la XIA au contexte spécifique de chaque application. Cette approche pose les bases de l'établissement d'une méthode d'évaluation éthique des mesures de la XIA pour les ingénieurs.

Précisons finalement que la matrice d'évaluation normative pourrait être utilisée à d'autres fins. Elle pourrait notamment être employée par des institutions publiques pour clarifier la portée de leurs recommandations. Comme nous l'avons mentionné dans la section 4.4, des institutions publiques comme la CEST proposent aux acteurs gouvernementaux des mesures concrètes visant à respecter différentes valeurs, allant de la transparence à la sécurité, en passant par le bien-être. Or, les avis de la CEST ne spécifient pas si les mesures recommandées sont relatives au contexte étudié, ou si elles sont universelles (les deux interprétations sont cohérentes avec la démarche de la CEST). À supposer que certaines des mesures proposées par la CEST soient relatives au contexte, le recours aux matrices d'évaluation permettrait à cette institution de mieux souligner l'adéquation entre (i) ses recommandations et (ii) les faits empiriques pertinents qu'elle prend en compte. D'une certaine façon, on pourrait dire que la matrice d'évaluation normative est un outil pour transmettre plus d'information aux décideurs quant aux éléments contextuels justifiant des mesures concrètes.

4.10 Remerciements

Nous remercions le Laboratoire d'ingénierie cognitive et sémantique (LiNCS) de l'École de technologie supérieure (ÉTS) pour son soutien, le Groupe de recherche interuniversitaire sur la normativité (GRIN) et le Centre de recherche informatique de Montréal (CRIM). Nous remercions également les évaluateurs anonymes pour leurs commentaires constructifs.

4.11 Conflits d'intérêts

Les auteurs déclarent qu'ils n'ont pas de conflit d'intérêt pour cette recherche.

CHAPITRE 5

MERGING ROLES AND EXPERTISE: REDEFINING STAKEHOLDER CHARACTERIZATION IN EXPLAINABLE ARTIFICIAL INTELLIGENCE

Camélia Raymond[a], Sylvie Ratté[b], Marc-Kevin Daoust[c]

[a, b, c] Département de Génie Logiciel et des TI, École de Technologie Supérieure
Département des Enseignements Généraux, École de Technologie Supérieure
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article publié dans « 2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON) », décembre 2024

5.1 Abstract

Explainable Artificial Intelligence (XAI) strives to make Artificial Intelligence Systems (AIS) more understandable, thus tackling the "black box" challenge. However, successful implementation requires precise identification of XAI requirements, made complex by the absence of universally accepted protocols. Given the importance of identifying stakeholders in this quest, this article proposes an innovative framework to characterize them. We compare and merge two predominant approaches: role-based and knowledge-based characterization. The result is a novel framework, segmenting knowledge into sub-categories while linking them to specific roles. This XAI Roles and Knowledge Framework offers a flexible methodology that can be adapted to the nuances of each XAI project. By providing a balance between specificity and generality, this tool aims to effectively guide the implementation of XAI while ensuring that the stakeholders' needs are taken into account. By using this approach, XAI projects benefit from a more precise identification of needs, leading to outcomes more closely aligned with user expectations and greater transparency in AI decisions.

5.2 Introduction

Artificial intelligence (AI) has evolved remarkably over the years, propelling significant advances in various fields. However, many of these AI systems (AIS) operate like "black

boxes", making their decisions often obscure and difficult for users to understand. Faced with this limitation, eXplainable Artificial Intelligence (XAI) has emerged as a promising solution, seeking to make AI decisions comprehensible. However, accurately identifying XAI needs is proving complex without clearly defined protocols. Although the literature recognizes the importance of stakeholder identification as a fundamental step in identifying XAI needs, existing protocols still need improvement. These are often perceived as too generic or excessively specific to particular use cases. This article seeks to fill this gap by proposing an innovative stakeholder characterization framework adapted to contemporary XAI requirements.

The innovation of our approach lies in our process of comparison and clarification. We have analyzed the relevance of exclusively role-based versus knowledge-based stakeholder frameworks. By exploring the strengths and weaknesses of each approach, we shed light on the debate by demonstrating the importance of merging these two perspectives.

This leads us to introduce two frameworks: (1) A revised framework for characterizing the roles of XAI stakeholders. This redesign enables better identification and characterization of stakeholders, thus offering optimal alignment of their responsibilities and expectations with the objectives of XAI projects. By integrating a regrouping structure, we facilitate the future prioritization of XAI-related needs, taking into account stakeholder expectations. (2) A framework of expertise, or knowledge framework, is also proposed. More nuanced than its predecessors, it segments knowledge into sub-categories and merges the notions of context and degree of knowledge to create a detailed map of expertise. The combined use of these two frameworks results in a scalable, flexible and adaptable tool, which we call the XAI Roles and Knowledge Framework.

5.3 Literature review

The difficulty of identifying and characterizing different stakeholders and their needs for explicability is a significant challenge. As recognized by Suresh et al. (2021), most research

focuses on one of two methods for classifying stakeholders: the first is based on stakeholders' expertise (knowledge), and the second is based on their role in the AIS. In these different frameworks, explicability needs and objectives are primarily defined by the category to which a user belongs. The following sections examine these frameworks' implementation and their advantages and difficulties.

5.3.1 Characterize stakeholders based on their expertise

Characterizing users or stakeholders according to their expertise in XAI represents an area primarily explored in the literature before the 2020s. This characterization is based on the principle that stakeholders' knowledge dictates their XAI needs.

Studies offer various approaches to this end. For example, Yu and Shi (2018) propose a classification of stakeholders based on their knowledge in the field of AI. They distinguish four levels of expertise: beginner, practitioner, developer and expert. In another study, Mohseni *et al*. (2021) identify three categories relevant to XAI: AI novices, AI experts and data experts. Also, Suresh *et al.* (2021) recommend a division based on three specific domains: the AIS field, the data field and the environmental field (environments where human-AI interaction occurs).

However, several problems emerge from these approaches. The most notable is that such characterizations need to sufficiently distinguish between stakeholders with similar expertise levels but different objectives. For example, suppose a hospital uses an XAI system to help doctors diagnose rare diseases. Two doctors use the system. Both are classified as medical "experts". The first doctor is a medical researcher who uses XAI to examine atypical cases, study new trends and gain an in-depth understanding of disease mechanisms. Detailed technical explanations are essential for him. Although equally knowledgeable, the second doctor uses XAI in consultation to explain diagnoses rapidly and clearly to patients. For the latter, an in-depth technical explanation is less valuable than a simplified, visual one that he can show the patient for better understanding. Their objectives with the system are distinct.

Consequently, such a characterization, while relevant for classifying expertise, needs to be sufficiently complete to enable XAI needs to be identified. It needs to take into account the variability of objectives among stakeholders. However, it remains helpful in defining how explanations should be presented and conveyed to ensure correct understanding. Current literature tends to neglect this form of stakeholder characterization. On the other hand, it is still relevant to ensure that explanations are provided in a format and vulgarization adapted to the knowledge of the targeted stakeholders.

5.3.2 Characterize stakeholders according to their roles

Characterizing users or stakeholders according to their functional role with AIS represents a significant line of research in the literature. This characterization is based on the principle that a person's role within an organization (or during a human-AI interaction) directly influences their needs regarding explainability.

Starting with the use case of the autonomous car, Hussain et al. (2021) put forward three critical roles for developing an AIS: engineers and scientists, ethicists, and end-users and consumers. Suresh et al. (2021) criticize this approach, pointing out that frameworks based solely on roles fail to segment the problem space in sufficient detail and modularity. They cite, for example, the case of clinical diagnostic models. In this context, the " consumers " category could encompass doctors and patients. Yet these two groups, despite their similar roles in model consumption, would require different explanations due to their distinct levels of medical expertise. Furthermore, it is surprising that these authors do not consider regulatory entities essential stakeholders.

Recognizing these limitations, recent work has sought to refine the characterization of stakeholders by role. In particular, Liao et al. (2021) have proposed a framework incorporating new categories such as "decision-affected parties." Their proposal segments stakeholders as fol-lows: model developers, business owners or directors, decision-makers, impacted groups, and finally, regulatory bodies. Similar frameworks have been suggested by Arya et al. (2019),

Langer et al. (2021) and Dhanorkar et al. (2021). For example, Dhanorkar et al. (2021) differentiate between data scientists and developers, whereas the proposal by Langer et al. (2021) and that by Liao et al. (2021) place them in the same stakeholder classification. Whereas the nuances be-tween these frameworks lie mainly in the nomenclature and precision of stakeholder segmentation, one crucial point remains: the distinction between stakeholders affected by the decision and direct users is widely recognized in recent literature.

One of the criticisms frequently observed when using a role-based stakeholder framework is that expertise within the stakeholder categories is often overlooked (Suresh et al., 2021). For example, a stakeholder framework that introduces "the development team" would not consider the notable differences between the role of data scientist and front-end developer. Several elements differ between these two stakeholders, such as the tasks they have to perform and the moment of interaction with the AIS in the development phases. We can, therefore, assume that these two roles have different needs for explainability and transparency. It has been noted that several of these frameworks are keen to present general stakeholders so that these can be used for all use cases. However, this is not an approach to be prioritized because of the loss of relevant information. Therefore, it would be wise to establish a detailed framework rather than a general one to distinguish the nuances between stakeholders and better identify their needs. Depending on the nature of the project, some stakeholders may be removed from the list. Nevertheless, the level of detail of the framework would enable a more significant amount of in-formation to be retained, which is necessary for identifying the various XAI needs.

Also, it is observed that within the same role, expertise can vary. Piorkowski et al. (2021) have noted this disparity, pointing out that the diverse expertise of team members within the same role often stems from their varied training and experience. A role-based framework does not allow these distinctions to be made. Although characterizing stakeholders according to role offers a valid first approximation of AIS explainability needs, and more specifically of their interactions with the AIS, more is needed. A more nuanced segmentation, considering both the role and the level of expertise of the various stakeholders, seems necessary to develop a framework adapted to real needs regarding AI explainability.

5.3.3 Conclusions and perspectives of the literature review

It is widely recognized that identifying the stakeholders of an AIS is crucial for a better understanding and identification of XAI needs. However, the field has yet to reach a consensus on a specific framework for stakeholder characterization. Indeed, two fundamental strategies are emerging: characterizing stakeholders according to their expertise OR characterizing stakeholders according to their roles. It is essential to emphasize the importance of this OR, as it suggests a choice between two distinct strategies.

Our literature analysis has shown that each approach has its limitations. Characterizing stakeholders according to their expertise ignores their intentions, whereas role-based characterization overlooks the diversity of knowledge within the same role group. Furthermore, even within these two strategies, there is no clearly defined framework.

Nevertheless, an interesting observation emerges: the flaws associated with one approach are often addressed by the strengths of the other. This paves the way for an innovative solution to the respective problems of each perspective. Rather than choosing one or the other, it would be wiser to consider the complementary use of these two frameworks.

- Characterizing stakeholders by their roles: this strategy captures stakeholders' intentions as well as their interactions with the AIS. This makes it possible to determine what information is needed for each stakeholder and at what point in the system's life cycle it is required.
- Characterize stakeholders by their expertise: this strategy focuses on how information is presented to different stakeholders to ensure rapid and effective assimilation, considering each group's skills and knowledge.

By combining these two approaches, we could design a framework that recognizes both the importance of stakeholders' roles and the need to tailor information according to their

expertise. This fusion could lead to better customization of explainable AI systems, optimizing their usefulness and relevance for diverse users. To ensure that the characterization of stakeholders by their roles facilitates the identification of AI needs, it is also recommended that roles be defined in as granular terms as possible, even if this means deleting specific roles when they do not apply to a project. Also, to increase the effectiveness of characterizing stakeholders by their expertise, it is essential to define the different types of expertise involved and the degree of expertise since this can vary (i.e. novice, intermediate, expert).

5.4 Roles and expertise: a combined perspective for Characterizing stakeholders

In this section, we present our innovative protocol, the result of an in-depth analysis of existing methodologies and emerging needs in the XAI field. Whereas previous approaches have often focused on either roles or expertise, our protocol adopts a merged approach, exploiting the strengths of each perspective. By combining roles and expertise, we aim to offer a more nuanced and comprehensive understanding of stakeholders to enable future advances in identifying XAI needs.

5.4.1 Stakeholder Role Characterization

We present a novel framework for characterizing stakeholders by role, integrating several crucial elements that need to be addressed or implemented in previous approaches.

- Stakeholders affected by the decision: It is essential to distinguish between direct users of AIS and individuals or groups affected by decisions arising from these systems. Their needs and concerns may differ.
- Regulatory bodies: Regulatory bodies play a crucial role in assessing the ethical and legal compliance of AI applications and should not be overlooked.
- Detailed segmentation: Rather than opting for an overly generalist framework, we have sought to establish a detailed framework that takes into account the specific nuances between stakeholders.

- Flexibility: While segmentation is crucial, it is also essential that the framework is flexible enough to adapt to different AIS use cases and contexts.

The study by Langer *et al.* (2021) proposes a decomposition of stakeholders into five categories, namely users, developers, deployers, parties affected by the AIS decision and regulators. Based on their model, we propose a more exhaustive list of roles. We have also chosen to incorporate a characterization of stakeholders taken from Eason (1989). This approach allows us to categorize three types of stakeholders: primary, secondary, and tertiary. Since this method of stakeholder identification is recognized and regularly used in software engineering, we hypothesize that it can be successfully applied to AIS. Additionally, this method makes it possible to prioritize the needs that come from these stakeholders, which is a concept that could be applied to the XAI field and thus support companies in prioritizing XAI needs. Following the logical order, the primary stakeholders have the highest priority, whereas the tertiary ones have the lowest.

Stakeholders are identified comprehensively, so businesses can use this model to determine the stakeholders relevant to their projects. Following this process, here are the stakeholders we have identified. Figure 5.1 presents an overview of the stakeholders and their relations to each other.

Primary stakeholders: The people or groups who use the AIS and are directly impacted by its positive or negative results (Eason, 1989). In software development, the needs and expectations of primary stakeholders are often the most critical factors in design.

1) Parties affected by the decision: Parties affected by the decision include all those impacted by the AIS (Langer et al., 2021). Certain aspects of their lives depend on the decision of an AIS, such as the refusal of a mortgage or the recommendation of medical treatment.

Secondary stakeholders: People who work with the results produced by the system (Eason, 1989). Although the needs and expectations of secondary stakeholders are less critical than primary stakeholders, they can be just as influential and make it challenging to implement a new AIS within a company if their needs are met (IEEE, 1984).

2) Users: Users include all those who take the recommendations issued by the AIS into consideration when making a decision (Langer et al., 2021). Some studies, such as Arrieta et al. (2021), refer to this category of stakeholders as domain experts.

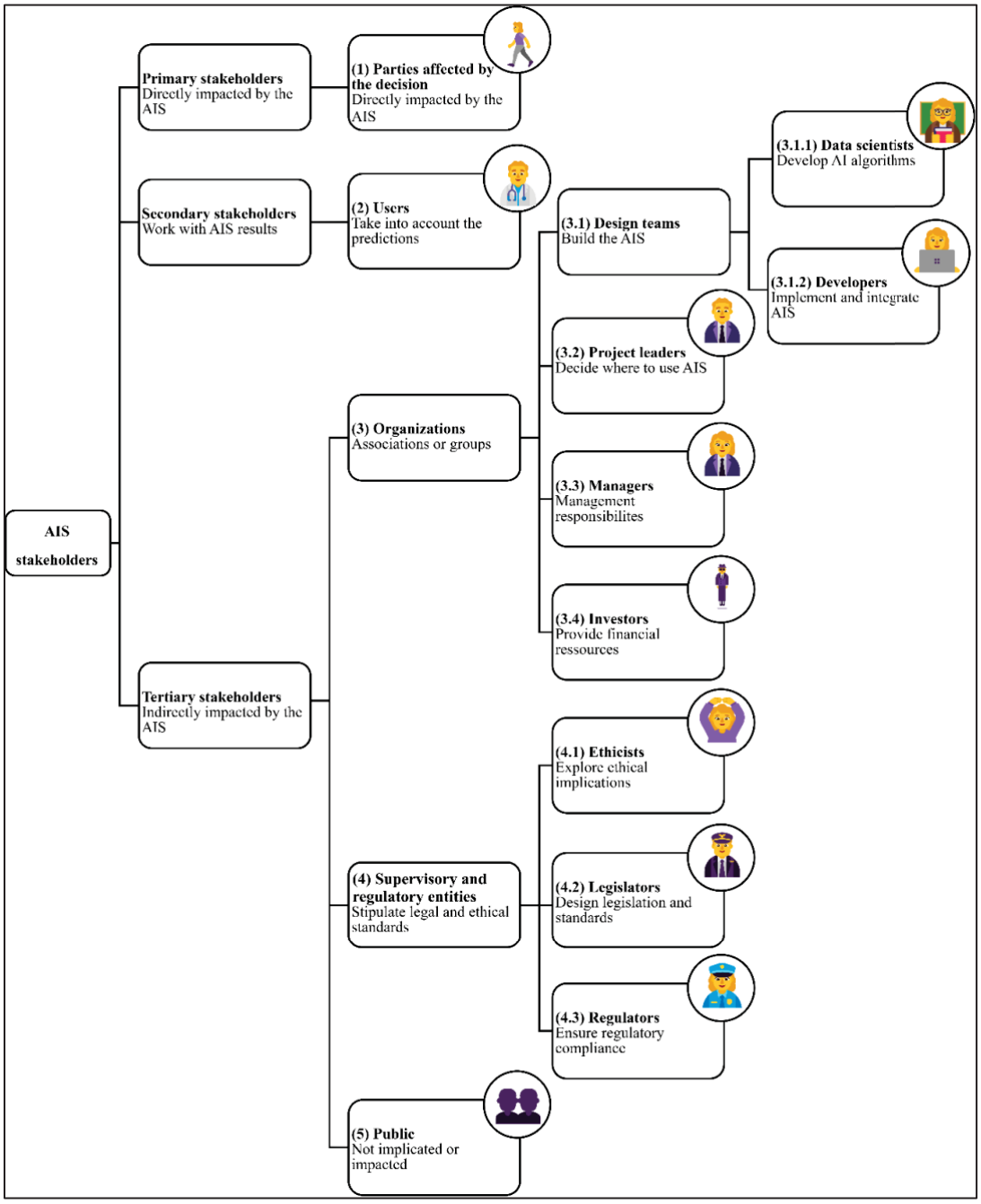


Figure 5.1 Role-based characterization of stakeholder

Tertiary stakeholders include people and groups affected more indirectly than secondary stakeholders (Eason, 1989). It is essential to involve tertiary stakeholders, as their opinions and perceptions can determine whether a project succeeds or fails.

3) Organization: Associations or groupings run by a moral or physical person to produce goods or services for the market.

3.1) Design teams: Design teams are the individuals who architect, program and build the AIS (Langer et al., 2021). Without them, AIS would not exist.

3.1.1) Data scientists: Data scientists develop AI algorithms and guide the understanding of AI systems from a scientific perspective (Suresh *et al.*, 2021).

3.1.2) Developers: Developers implement AI systems and integrate them into the IT infrastructure (Suresh et al., 2021).

3.2) Project leaders: Project leaders are the people who decide where to employ specific systems (Langer et al., 2021). They manage the overall application domain in which AI systems operate. They do not necessarily interact directly with the AIS (Suresh et al., 2021). Their decisions influence several other classes of stakeholders. They often act as intermediaries between several stakeholders.

3.3) Managers: Managers manage a business, service, or administration (Weller, 2019).

3.4) Investors: Investors enable the development and use of AI systems by providing the necessary financial resources to organizations (Suresh et al., 2021).

4) Supervisory and regulatory entities: Supervisory and regulatory entities include all regulators who stipulate legal and ethical standards for the general use, deployment and development of AIS (Langer et al., 2021).

4.1) Ethicists: Ethicists explore AIS's ethical, social and philosophical implications (Suresh et al., 2021).

4.2) Legislators: Legislators design legislation and standards and are responsible for the legal regulation of AI systems (Suresh et al., 2021).

4.3) Regulators: Regulators ensure compliance with the legal regulation of AIS (Suresh et al., 2021).

5) Public: The public includes all people who are not involved in or impacted by the project but are aware of its existence.

5.4.2 Stakeholder knowledge characterization

We present a novel framework for characterizing stakeholders by knowledge, integrating several crucial elements that need to be addressed or implemented in previous approaches. Suresh et al. (2021) have established three knowledge contexts. These are AIS knowledge, data knowledge and environment knowledge. We add a layer of precision for each of these contexts by adding sub-categories of expertise to create a more complex model better adapted to

business reality. In addition, we determine the degree of knowledge for each knowledge context.

5.4.2.1 Knowledge contexts

AIS knowledge (C1) represents the knowledge required to research, develop, operate or deploy machine learning models Suresh et al. (2021). We can divide AIS knowledge into two subgroups: technical knowledge in artificial intelligence (C1.a) and AIS integration knowledge (C1.b). Technical knowledge in artificial intelligence consists of the knowledge required to research and develop machine learning models - for example, learning algorithms, algorithm optimization and model evaluation. AIS integration knowledge refers to all the knowledge required to operate and deploy AIS - for instance, deploying production models, risk management, maintenance, and updating.

Data knowledge (C2) represents the knowledge required to collect, organize, analyze and communicate the data with which the model has been trained and makes decisions Suresh et al. (2021). We can further divide data knowledge into two subgroups: data integration knowledge (C2.a) and data domain expertise knowledge (C2.b). Data integration knowledge consists of the knowledge required to collect and store data - for example, data collection, storage, security and confidentiality. Data domain expertise knowledge refers to all knowledge of the theory related to the data domain of expertise and related technologies. For example, this knowledge could be domain-specific theories, knowledge of domain trends and developments, and data interpretation and validation skills. More specifically, we could appoint domain experts who can interpret a lung X-ray to detect cancer.

Environmental knowledge (C3) represents knowledge of the environments in which human-AI interaction can occur Suresh et al. (2021). We can further divide this knowledge context into two subgroups: knowledge of the physical environment (C3.a) and the socio-cultural environment (C3.b). Knowledge of the physical environment consists of the geographical locations, or places, where the AIS will be put into operation and used. For example, physical

information about a house, a bank, a courthouse or a doctor's clinic. Knowledge of the socio-cultural environment is knowledge of all the events, facts and phenomena (relating to social, cultural and economic aspects) observed in the physical environment under study. For example, historical biases, culture, values and social classes.

5.4.2.2 Knowledge degree

The knowledge level could be interpreted as the degree of comfort with which an actor can dis-cuss a knowledge context. We propose to distinguish three degrees of knowledge: advanced knowledge (D1), basic knowledge (D2) and no knowledge (D3).

An actor with advanced knowledge of a knowledge context possesses formal and instrumental knowledge. Formal knowledge comprises an understanding of codified theories embodied in texts or diagrams such as those found in textbooks and is acquired during a prolonged educational process (Eraut, 2010). Instrumental knowledge is an understanding of how to "apply" formal knowledge. It is embodied using tools or other instruments and learned through demonstration and practice (Eraut, 2010).

An actor with basic knowledge of a knowledge context has personal knowledge of that context. Personal knowledge describes information entirely embodied by individuals and acquired through their participation in specific domains. This type of knowledge is difficult to codify (Eraut, 2010).

Actors with no context knowledge means they have no formal, instrumental or personal knowledge. They would, therefore, never have come into contact with this type of context in the past.

5.4.3 Combining roles and knowledge

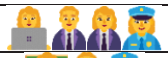
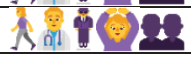
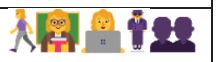
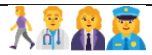
Having defined the frameworks for stakeholder knowledge and roles, it's time to merge them. A crucial nuance in this merging is that, in situations where a specific role might have higher-

level knowledge, we have opted for a lower-level approach. This decision is based on the belief that offering simplified explanations is better than adopting a complexity that might be difficult to grasp.

It is also essential to stress that the link between roles and knowledge must be reviewed at the start of each project. Although our framework presents a typical association observed in most projects, variations may occur according to the specificities of each initiative, allowing specific roles to have adapted or even more in-depth knowledge.

Table 5.1 presents the merged role-based and expertise-based framework.

Tableau 5.1 Context and degree of knowledge of AIS stakeholders

	(D1) Advance	**(D2) Basic**	**(D3) None**
(C1) AIS knowledge			
(C1.a) Technical AI knowledge			
(C1.b) AIS integrations knowledge			
(C2) Data knowledge			
(C2.a) Data integration knowledge			
(C2.b) Data expertise knowledge			
(C3) Environment knowledge			
(C3.a) Physical environment knowledge			
(C3.b) Socio-cultural environment knowledge			

This innovative framework merges role-based and expertise-based approaches to comprehensively appreciate the various players involved. It enables stakeholders to be segmented into dis-tinct categories, considering both their functional role and their specific level of expertise. This method represents an essential first step in developing best practices for defining XAI requirements by facilitating the precise identification of each stakeholder's needs. The information gathered can be used to determine each stakeholder's information requirements and needs precisely, as well as the level of vulgarization required. Therefore, this

framework serves as a guide for identifying and a basis for defining communication and information needs in XAI projects.

5.5 Considerations for making the most of this innovative framework

Our framework offers unique flexibility, enabling it to be adapted to various contexts and situations. It has been designed to be remarkably adaptable downwards by excluding users present in the initial framework but absent from the project context. Imagine an AIS that does not use personal data or produce predictions affecting an individual; in such an instance, no stakeholder may be affected by the decision. Similarly, the design team might rely solely on a data scientist when producing a simple proof of concept. Consequently, the developer would not be present in the outcome of stakeholder identification.

However, exceptional situations may require roles to be added to the framework. For example, if a product is used by users performing distinct tasks, as demonstrated in the previous medical example, each would have his or her intentions. This distinction could be translated into a segmentation into "Emergency Center User" and "Research Center User," making it easier to identify specific XAI needs.

Beyond these adaptations, it’s crucial to associate each role with specific people, whether identified by name or grouped by function. The same person can take on several roles. For example, in a fully automated AIS context, the user could also be the stakeholder affected by the decision. Similarly, one person within an organization could be a data scientist and a developer.

To make the most of this framework, we recommend reviewing and confirming these role associations at the launch of each new project. Although our tool proposes a typical association observed in many projects, the specificities of each initiative may require adjustments. In this way, the framework becomes a living tool designed to evolve according to stakeholders’ needs and feedback, guaranteeing its relevance and effectiveness for all XAI projects.

5.6 Conclusion and prospects for the XAI Roles and Knowledge Framework

In this article, we have navigated the complexities of identifying and characterizing stakeholders in XAI. Instead of traditional approaches, often reduced to silos of perspectives, we have introduced an innovative approach that merges the concepts of roles and knowledge. Our XAI Roles and Knowledge Framework has established itself as a promising reference for deciphering and navigating the complexities of the XAI domain. Recognizing the importance of Eason's (1989) characterizations makes our proposal more robust and facilitates the challenging task of prioritizing XAI requirements.

As established in the majority of scientific articles defining the needs and requirements of XAI projects, the definition of stakeholders is an essential first step in developing best practices for defining XAI requirements. It is becoming increasingly clear that the role-based framework should guide the characterization of interactions between XAIs and stakeholders. It is crucial to determine precisely what information each stakeholder seeks and their specific intentions. Once this determination has been achieved, the focus can shift to the nature and quantity of information to be shared with them. Furthermore, the knowledge framework can be used to refine the communication strategy, ensuring that information is presented in a manner that is both relevant and accessible based on the stakeholders' prior knowledge. Together, these two axes converge towards a vision where explanations are meticulously tailored, ensuring optimal transmission of information.

However, like any field in constant evolution, XAI will continue to transform, with it, the stakeholders' needs and requirements. While flexible and progressive, our framework will need to be reassessed and adapted to keep pace with technological advances and feedback from the industry.

Ultimately, we're convinced that our approach provides a solid foundation for improving the effectiveness and relevance of XAI initiatives. By redefining how we perceive and integrate

stakeholders, we are paving the way for AI projects that are more understandable, transparent, and, above all, better aligned with the needs of all stakeholders.

5.7 Acknowledgements

I sincerely appreciate the École de Technologie Supérieure (ÉTS) for their academic support and resources, which have been instrumental to this research journey. My gratitude also goes to the Centre de Recherche Informatique de Montréal (CRIM), where collaboration and the research culture contributed to my work. Thanks also to the anonymous reviewers, whose copious suggestions significantly improved the presentation of this paper.

CHAPITRE 6

UNE APPROCHE MÉTHODOLOGIQUE DE L'IDENTIFICATION DES EXIGENCES, DES CONTRAINTES ET DES FONCTIONNALITÉS DE LA XIA : EXPLAINABILITY AI REQUIREMENTS SPECIFICATION

6.1 Introduction

L'avènement de l'intelligence artificielle (IA) a révolutionné de nombreux domaines, allant de l'industrie à la santé, en passant par l'éducation et la finance. Toutefois, cette intégration massive de l'IA soulève d'importantes préoccupations éthiques et de gouvernance, au cœur desquelles se trouve la nécessité d'une IA explicable (XIA). Dans un contexte où les algorithmes peuvent influencer des décisions critiques touchant à la vie des individus, que ce soit à travers des diagnostics médicaux (par exemple, Caruccio et al., 2024), des opérations financières (par exemple, Ahmadi, 2024), ou même des jugements judiciaires (par exemple, Atkinson et al., 2020), il devient crucial que ces systèmes soient non seulement performants mais aussi transparents et compréhensibles par tous. La XIA cherche à répondre à cette problématique, en s'assurant que les décisions prises par les IA puissent être expliquées et comprises (Chromik et Butz, 2021), facilitant ainsi la création d'une confiance des humains envers l'IA (Ali, 2023). C'est dans les domaines critiques, où les enjeux sont particulièrement élevés, que l'urgence de développer et d'adopter des technologies d'IA explicable se fait le plus sentir.

Dans ce contexte, nous proposons, mettons en œuvre et validons une méthodologie originale d'identification des exigences, des contraintes et des fonctionnalités de la XIA, un domaine où peu de réponses ont été amenées et acceptées dans la communauté scientifique, mais essentiel pour la confiance et la transparence des systèmes d'IA.

La compréhensibilité n'est pas seulement une question de convivialité mais aussi un impératif éthique et légal. Plusieurs initiatives et cadres réglementaires à travers le monde commencent à souligner cette nécessité. Par exemple, passé en 2016, le *Règlement Général sur la Protection*

des Données (RGPD) de l'Union Européenne stipule le droit à l'explication, obligeant les organisations à fournir des informations claires sur les logiques sous-jacentes, ainsi que sur l'importance et les conséquences prévues du traitement des données par des systèmes automatisés. Aux États-Unis, le projet de loi *Algorithmic Accountability Act* (2022) propose des évaluations d'impact sur les algorithmes pour les systèmes à haut risque, exigeant une transparence et une explicabilité accrues des décisions automatisées. Ces exemples législatifs transatlantique illustrent l'émergence d'une conscience mondiale autour de l'importance de rendre les IA non seulement performantes mais également éthiques et transparentes.

Pour les concepteurs de systèmes d'IA cherchant à se conformer aux lois sur la XIA, la tâche est particulièrement ardue en raison de l'absence de bonnes pratiques clairement établies dans la littérature. Il est difficile de déterminer quels outils répondent à quels besoins et à quelles lois, et de surcroît, les besoins eux-mêmes sont souvent mal définis. Par exemple, dans le secteur de la santé, un modèle d'IA destiné au diagnostic médical exige une explicabilité pour permettre aux médecins d'interpréter ses recommandations. Dans le domaine financier, l'explicabilité est cruciale pour garantir la transparence et l'équité des décisions automatisées de crédit. Ces exigences, toutes aussi importantes les unes que les autres, sont très variées. De plus, le cadre législatif, les exigences éthiques et techniques autour de l'IA évoluent rapidement, rendant indispensable une approche structurée pour intégrer l'explicabilité au cœur des projets d'IA.

Nous avons développé et validé l'outil *Explainable AI Requirements Specification* afin de répondre à ces défis. Inspiré par les normes de spécifications de requis logiciels de l'Institute of Electrical and Electronics Engineers (IEEE), cet outil vise à fournir un cadre méthodologique pour spécifier de manière structurée les exigences en matière d'explicabilité dans les systèmes d'IA, les contraintes ainsi que les fonctionnalités. L'objectif est de guider les concepteurs de systèmes d'IA à travers des exigences légales, éthiques et techniques, en éclaircissant les aspects clés nécessaires à la création d'une IA explicable.

Dans la première partie de cet article (section 6.3), nous examinons les défis et les opportunités présentés par la XIA, ce qui nous amène à souligner l'importance de créer une approche méthodologique pour définir les besoins. Ensuite, nous explorons la démarche employée pour développer l'outil *Explainable AI Requirements Specification*, en discutant de ses principes de conception, de l'approche multidimensionnelle utilisée ainsi que la méthodologie de validation dans des contextes industriels (section 6.4). Puis, notre article décrit l'architecture du *Explainable AI Requirements Specification* (section 6.5). Cette description vise à fournir une compréhension de la manière dont l'outil peut formuler les fonctionnalités de la XIA, et leurs intégrations dans le développement des systèmes d'IA. Nous décrivons également l'utilité du *Explainable AI Requirements Specification Template for Common AI Projects* (section 6.6), créé dans le but de faciliter la rédaction du *Explainable AI Requirements Specification*. Puis l'article se conclut sur une évaluation et les perspectives de l'évolution de ses outils (section 6.7).

6.2 Les défis et opportunités de la XIA

La XIA se présente comme une réponse à la complexité croissante des systèmes d'IA, notamment lorsqu'ils sont déployés dans des domaines critiques, mais elle apporte aussi un ensemble unique de défis. Ces défis multidimensionnels sont à la fois techniques, éthiques, et légaux. Ils nécessitent une attention particulière pour garantir le développement de systèmes d'IA performants et compréhensibles. La XIA émerge alors comme un vecteur de confiance.

Une IA explicable peut être perçue comme un morceau du casse-tête que représente l'IA de confiance. Une XIA permet, conjointement avec d'autres principes, de répondre au manque de confiance en l'IA. Selon Les lignes directrices en matière d'éthique pour une IA digne de confiance (Commission européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019), le concept d'explicabilité, dans le domaine de l'IA, se définit comme suit :

« L'explicabilité concerne la capacité d'expliquer à la fois les processus techniques d'un système d'intelligence artificielle et les décisions humaines qui s'y rapportent (par exemple, domaines d'application d'un système d'intelligence artificielle). L'explicabilité technique suppose que les décisions prises par un système d'intelligence artificielle peuvent être comprises et retracées par des êtres humains. [...] Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur) » (Commission Européenne et Direction générale des réseaux de communication, du contenu et des technologies, 2019).

De cette définition, nous pouvons identifier quatre concepts clés. Le premier est « le processus technique » qui, en son sens, pourrait faire référence à la logique de l'algorithme et les stratégies de conception du système d'un point de vue IA, mais aussi du point de vue plus général des technologies de l'information. Le deuxième est « les décisions humaines qui s'y rapportent ». Ici, l'on fait référence à une action qui est réalisée à la suite de la prise d'information présentée dans l'explication. La définition mentionne aussi que les explications doivent être « présentées en temps opportun » et « adaptée à l'expertise de la partie prenante ».

Dans la pratique, à quoi ressemble réellement une IA explicable ? Quelques outils ont été développés et sont couramment utilisés sous la terminologie de la XIA. Les deux outils les plus communs dans la littérature sont *Local Interpretable Model-Agnostic Explanations* (LIME) et *SHapley Additive exPlanations* (SHAP) (Slack et al., 2020). La figure 6.1 et 6.2 montrent des tableaux de bord développés à l'aide de ses outils.

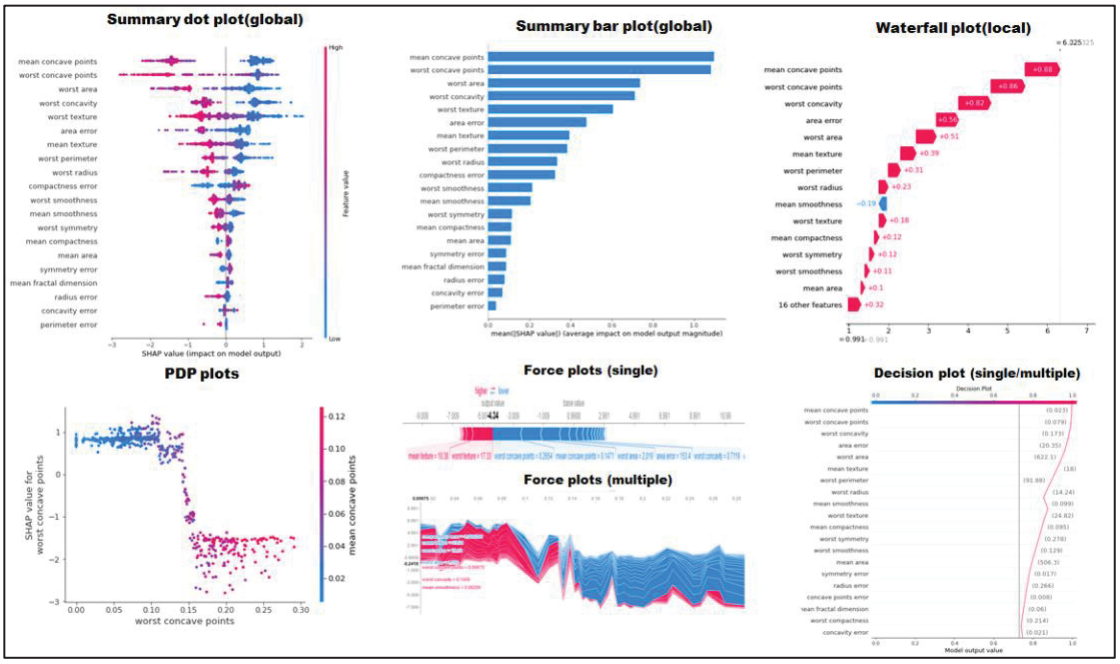


Figure 6.1 Tableau de bord qui présentent de explications créées avec l'outil SHapley Additive exPlanations,

Tirée de Bhatnagar (2021)

Pour les concepteurs d'IA, ces approches sont importantes et permettent d'enrichir la compréhension des systèmes d'IA qu'ils créent. Cependant, ces outils ne sont pas toujours pertinents. Par exemple, il est difficile d'imaginer qu'un médecin, dans une salle d'urgence, possédant que quelques secondes et peu de connaissances en IA, puisse tirer des informations pertinentes de ces explications pour émettre une recommandation sur un choix de traitement.

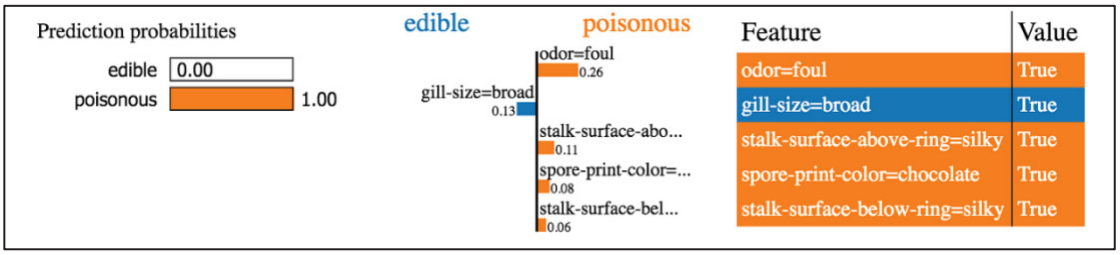


Figure 6.2 Tableau de bord qui présentent de explications créées avec l'outil Local Interpretable Model-Agnostic Explanations,

Tirée de Yasaini (2023)

Lorsque l'on se ramène à la définition de la XIA, il apparaît que les outils disponibles en pratique ne répondent pas aux exigences. Nous soulignons que les visualisations proposées ne sont pas toujours pertinentes par rapport aux décisions humaines qui s'y rapportent, comme illustré dans l'exemple médical précédent. De plus, ces outils ne précisent pas pour quel moment les explications ont été conçues et ne sont pas adaptées aux niveaux d'expertise de toutes les parties prenantes concernées.

Pourquoi alors est-il si complexe de créer des outils qui répondent aux besoins de l'industrie en XIA ? Au cœur des défis se trouve un manque flagrant de normes, de méthodologies et de protocoles assurant le développement efficace, dirigé et guidé d'IA explicables. Plusieurs facteurs contextuels et techniques exacerbent cette carence de méthodologies en XIA.

Premièrement, nos recherches montrent qu'il y a des lacunes dans la clarification de terminologies dans la législation portant sur la XIA. Par exemple, au Québec, la Loi 25 exige que les entreprises privées, qui exploitent des systèmes complètement automatisés et qui utilisent des renseignements personnels, fournissent, à la demande de la personne concernée, les raisons, ainsi que les principaux facteurs et paramètres, ayant mené à la décision (Gouvernement du Québec, 2021). L'utilisation d'un vocabulaire imprécis tel que « des raisons, ainsi que des principaux facteurs et paramètres » laisse une grande place à l'interprétation. Ses lacunes génèrent de la confusion, notamment pour les ingénieurs spécialisés en IA souhaitant se conformer aux normes, sans vraiment comprendre nettement les attentes. Cette ambivalence crée un terrain propice à des utilisations incorrectes et une difficulté accrue de création de méthodologies qui assurent le respect de la loi en XIA.

Deuxièmement, l'absence de méthodologie pour développer des mesures concrètes adaptées au contexte de chaque entreprise qui respectent leurs engagements éthiques. Au Canada, en date du 18 mai 2024, La déclaration de Montréal pour un développement responsable de l'intelligence artificielle (DDIRIA) (Université de Montréal, 2018) a été signée par 173 organismes et entreprises canadiennes, dont plusieurs organismes fédéraux, centres de

recherches et entreprises privées. La DDIRIA est une œuvre collective *qui a pour objectif de mettre le développement de l'IA au service du bien-être de tous et chacun, et d'orienter le changement social en élaborant des recommandations ayant une forte légitimité démocratique* (Université de Montréal, 2018). Cette adhésion reflète une volonté de développer des systèmes d'IA qui sont non seulement éthiques mais aussi transparents et explicables (principes clés présentés dans la DDIRIA). Bien que l'intention soit claire, l'industrie se heurte à des difficultés significatives dans l'implémentation pratique de ces principes éthiques dans le développement des systèmes d'IA. Pour atteindre cet objectif, les entreprises doivent développer, en fonction de leurs contextes, des mesures concrètes personnalisées, et être capables de valider leur mise en œuvre (Floridi, 2019). Actuellement, la littérature proposant une méthodologie qui permettrait de relier des mesures concrètes à des principes éthiques est limitée.

Troisièmement, le manque de justification de la robustesse de la XIA. Pour parvenir à une XIA robuste, plusieurs composants manquent encore. Actuellement, il existe peu de protocoles pour identifier les besoins des parties prenantes en XIA, ce qui limite la création et la priorisation des métriques d'évaluation pertinentes. Il est donc possible de créer des explications, mais nous n'avons pas encore de méthodologie claire qui nous permettent de valider si ses explications sont exactes, complètes, consistantes, de complexité ajustée à l'audience ou encore cohérentes. Toutes des caractéristiques identifiées dans la littérature comme étant nécessaires à la création d'une bonne explication, selon la revue de Nauta et al. (2023). De plus, plusieurs caractéristiques d'une bonne explication pourraient se contredire (Neely et al. 2021). Par exemple, un niveau de complexité moindre et la complétion d'une explication. La création d'une méthodologie d'identification des besoins en XIA pourrait supporter la création de métriques qui permettent d'évaluer ses facteurs, en plus de les prioriser suivant une priorisation des besoins du projet.

Quatrièmement, la diversité de définition des cycles de vie des systèmes d'IA. Les entreprises qui créent des systèmes d'IA décrivent de manière variée les cycles de vie de ces systèmes. Par exemple, certaines entreprises utilisent une approche de développement appelée Model-

Cenetric AI[1] et d'autres utilisent une approche nommée Data-Centric AI[2]. Aussi, certaines entreprises n'utilisent pas d'approche prédéfinie. La définition du cycle de vie est importante puisque cela permet de définir le contexte d'utilisation de l'IA, qui à son tour favorise l'évaluation de l'efficacité de mesures concrètes en XIA ainsi que l'identification des besoins (Dhanorkar et al., 2021). Les variations possibles dans la définition des cycles de vie des systèmes d'IA rendent difficile la conception d'une approche méthodique et structurée, applicable à toutes les entreprises et secteurs, pour la création de systèmes d'IA explicables.

Cinquièmement, des lacunes dans l'identification des parties prenantes. Il est largement reconnu que l'identification des parties prenantes d'un système d'IA est cruciale pour une meilleure compréhension et identification des besoins en XIA (Dhanorkar et al., 2021). Leur importance est par ailleurs mentionnée directement dans la définition de la XIA. Toutefois, il n'y a pas encore de consensus à propos d'un cadre précis pour la caractérisation des parties prenantes. Deux tactiques majeures émergent : caractériser les parties prenantes en fonction de leurs expertises - par exemple, Yu et Shi (2018), Mohenseni et al. (2021) - OU caractériser les parties prenantes en fonction de leurs rôles – par exemple, Langer et al. (2021), Dhanorkar et al. (2021), Liao et al. (2021). Ce manque de consensus complexifie la création d'une méthodologie d'identification des besoins en XIA puisque ceux-ci découlent des parties prenantes du système.

Ces éléments mettent en évidence que le manque de méthodologies et la réalité des outils limités disponibles constituent un risque majeur pour le succès des projets en IA explicables.

---

[1] L'approche Model-Centric AI (MCAI) représente une méthodologie dominante dans le développement de l'intelligence artificielle, mettant l'accent sur l'optimisation des modèles en termes de précision, de performance et d'efficacité computationnelle. Cette approche a façonné la manière dont l'IA a été abordée dans le domaine de la recherche et de l'industrie pendant des décennies (Hamid, 2022).

[2] Dans les dernières années, le Data-Centric AI (DCAI) a émergé comme un paradigme révolutionnaire, soulignant l'importance capitale de la qualité et de l'ingénierie des données dans la création de systèmes d'IA performants (Majeed et Hwang, 2023). À la différence de l'approche MCAI, qui se concentre principalement sur le perfectionnement du modèle d'IA lui-même, le DCAI met l'emphase sur l'amélioration des données, reconnaissant ainsi que la qualité des données est un facteur déterminant pour la performance d'un système d'IA.

Tant qu'une approche méthodique et acceptée par la communauté scientifique ne sera pas accessible, cela constitue un risque majeur pour le succès des projets en XIA.

6.3 Démarche de développement de l'outil *Explainable AI Requirements Specifications*

Nous avons développé l'outil *Explainable AI Requirements Specification (XAIRS)* en réponse au manque de méthodologie dans le développement des technologies de la XIA. Le XAIRS est un cadre méthodologique pour spécifier de manière structurée les exigences en matière d'explicabilité dans les systèmes d'IA, les contraintes ainsi que les fonctionnalités.

La démarche employée pour développer l'outil *Explainable AI Requirements Specification* (XAIRS) se distingue par son caractère multidisciplinaire, sa capacité d'adaptation aux standards de l'industrie et sa flexibilité face aux différents contextes et domaines d'utilisation (exemple : médical, finance et juridique).

Avant de vous présenter le XAIRS (section 6.5), nous introduisons la démarche pour le développer. L'objectif de cette section est de décrire le processus de développement de l'outil en y présentant les grandes phases (section 6.4.1) et ensuite, en détaillant chacune de celles-ci (section 6.4.2, 6.4.3 et 6.4.4).

6.3.1 Vue globale du processus de développement du *Explainable AI Requirements Specification*

Le processus de développement du XAIRS a été réalisé en quatre étapes clés : l'analyse de la littérature en XIA, la conception et le prototypage de l'outil, la phase d'amélioration et de validation du XAIRS par des relectures d'experts, et finalement une évaluation quantitative de l'utilité de l'outil au travers de tests réalisés en industrie.

La première étape était l'analyse de la littérature pour identifier les besoins et les défis associés à la XIA. Cela inclut des consultations avec des parties prenantes clés, experts dans leurs

domaines respectifs et une revue de littérature approfondie pour comprendre les différentes dimensions de la confiance, l'impact de la transparence et plus particulièrement de l'explicabilité dans les systèmes d'IA. Les thématiques suivantes ont été étudiées : les systèmes d'IA de confiance, l'explicabilité pour augmenter la confiance, les systèmes d'IA explicables licites, éthiques, robustes, le cycle de vie ainsi que les parties prenantes des systèmes d'IA explicable. Une synthèse de cette revue a été présentée à la section 6.3. La revue complète est présentée au chapitre 1.

Après l'analyse, l'étape suivante a été la conception et le développement du XAIRS. Cette phase a impliqué la création du prototype initial. La section 6.4.2 décrit les choix de conceptions réalisés lors de cette phase.

Le XAIRS a ensuite été soumis à plusieurs cycles de validation et d'amélioration. Des experts en IA, des juristes et des professionnels de l'éthique ont été impliqués lors de relecture pour s'assurer que l'outil répondait aux exigences techniques, éthiques et légales. La section 6.4.3 décrit notre approche multidisciplinaire qui différencie notre recherche de celles souvent observées dans la littérature, pour cette étape clé de validation.

Finalement, l'outil a été soumis à plusieurs tests en industrie, notamment dans 3 projets, tous provenant de secteurs d'activités différents. Les secteurs d'activités sont la défense, le transport et la santé, tous des secteurs d'activités à fort impact sur le public. Ces phases de tests ont permis d'améliorer, de manière itérative, le XAIRS et de récolter des retours quantitatifs sur sa pertinence. Ces itérations sont décrites à la 6.4.4.

6.3.2 Choix de conception du *Explainable AI Requirements Specification*

L'élaboration de l'outil XAIRS repose sur une approche rigoureuse, visant à intégrer efficacement l'explicabilité dans les systèmes d'IA. Elle s'appuie sur une fondation théorique qui incorpore les dernières recherches techniques et pratiques en IA, l'éthique de l'IA, ainsi que

les normes légales. Afin d'assurer un développement cohérent du XAIRS, nous avons choisi certains principes de conception que nous avons suivi tout au long de l'élaboration de l'outil :

1) Clarté et intelligibilité : L'outil est conçu pour être accessible et facilement compréhensible, permettant aux ingénieurs logiciels, quelle que soit leur spécialisation sectorielle, de suivre et d'appliquer ses directives.
2) Conformité aux normes éthiques et légales : L'outil prend en compte les exigences légales et éthiques actuelles en matière d'IA, contribuant ainsi au déploiement de systèmes respectant ces contraintes.
3) Adaptabilité et flexibilité : La conception permet une adaptation aux différents contextes sectoriels des projets d'IA, offrant une structure flexible et robuste.

Ces principes ont été appliqués et ont affecté le format final de l'outil. Par exemple, le principe de clarté et d'intelligibilité nous a amené à produire des exemples, directement inclus dans l'outil, afin de guider les ingénieurs logiciels qui utiliserait cet outil, sans avoir des connaissances approfondies dans le domaine de la XIA. Le principe de conformité aux normes éthiques et légales nous a amené à créer des sections spécifiques, dédiées à ses thèmes. Le principe d'adaptabilité et de flexibilité est présent dans toutes les sections, où aucune d'entre elles n'est spécifique à un domaine d'affaire, mais présente plutôt une méthodologie pour spécifier la XIA au domaine de leur choix.

S'inspirant du *Software Specification Requirements* (SRS), outil de développement de normes de spécifications de logiciels créé par l'IEEE (1984), le XAIRS est structuré de manière à être à la fois familier aux ingénieurs en logiciel et adapté aux spécificités de la XIA. Cette intégration assure que l'outil est à la fois pratique et conforme aux standards de l'industrie. Le choix de s'inspirer du SRS, initialement dans l'intention de faciliter son adoption, s'est avéré judicieux tant pour simplifier la phase de conception du XAIRS que pour garantir un standard de qualité.

Le XAIRS incorpore des sections similaires à celles présentes dans le SRS, telles que l'introduction, la description générale et les contraintes opérationnelles. Cette structure, familière aux ingénieurs habitués à rédiger des SRS et à l'industrie, facilite l'adoption et l'utilisation de l'outil XAIRS. Toutefois, l'outil se distingue par l'intégration de sections spécifiquement dédiées à l'explicabilité de l'IA, notamment l'évaluation des besoins en

informations des parties prenantes et les fonctionnalités de la XIA. Ces sections sont décrites à la section 6.5.

6.3.3 Intégration de l'approche interdisciplinaire pour l'amélioration graduelle du *Explainable AI Requirements Specification*

Dans le développement de l'outil XAIRS, une approche interdisciplinaire a été adoptée pour capturer la complexité et la diversité des défis posés par la XIA. Cette démarche reflète la conviction que la collaboration entre divers domaines d'expertise est essentielle pour créer des solutions d'IA.

Pour construire cet outil, nous avons eu recours à des personnes exerçant une multitude de professions, telles que : des ingénieurs, des développeurs logiciels, des scientifiques des données, des experts en traitement automatisé des langues naturelles (TALN), des experts en vision par ordinateur, des gestionnaires de projet en IA, une ex-juriste, un avocat et des experts en éthiques. Les contributions de chacun ont été cruciales pour développer un outil complet qui soit à la fois techniquement robuste et socialement responsable.

L'outil a aussi été conçu pour être adaptable à différents contextes d'utilisation. Que ce soit dans des environnements réglementés comme la santé ou la finance, ou dans des applications plus générales, l'outil peut être personnalisé pour répondre aux exigences spécifiques de chaque domaine. Afin de valider son utilité multi-sectoriel et contextuel, obtenu par notre approche interdisciplinaire, le XAIRS a pu être testé et validé dans des projets industriels provenant de divers secteurs.

De plus, le XAIRS peut être utilisé dans des contextes de recherches académiques et industriels. Il offre un cadre pour intégrer l'explicabilité dès les premières étapes de la conception des systèmes d'IA, même si son industrialisation n'est pas (encore) envisagée. L'outil possède l'avantage d'avoir été développé dans un cadre académique et testé dans un cadre industriel, ce qui lui permet d'être versatile.

6.3.4 Protocole de test du XAIRS dans des contextes industriels réels et variés

La phase de tests de l'outil XAIRS a été cruciale pour valider son efficacité et son adaptabilité dans des environnements réels et variés. Pour ce faire, nous avons utilisé l'outil dans trois secteurs d'activité distincts : la défense, le transport et la santé. Chacun de ces secteurs présente des défis uniques en matière de transparence et d'explicabilité de l'IA, ce qui nous a permis d'évaluer la robustesse et la flexibilité de l'outil. Les tests dans ces différents contextes industriels ont permis de valider la polyvalence et l'efficacité du XAIRS. Chaque phase de test a été suivie d'itérations, basées sur les retours des utilisateurs et des experts. Ces itérations ont conduit à des améliorations significatives de l'outil, notamment en termes de convivialité, de précision des spécifications et d'adaptabilité aux normes sectorielles.

Dans le secteur de la défense, l'explicabilité des systèmes d'IA est cruciale pour assurer la sécurité et la confiance. Le XAIRS a été testé dans un projet visant à améliorer la détection des menaces. Cette phase de test a été particulièrement utile pour améliorer l'identification des parties prenantes dans un projet en XIA et leurs caractéristiques, notamment leurs expertises. Il a été observé un contraste notable entre les types de connaissances des utilisateurs, soit aucune connaissance en IA, mais des connaissances très poussées dans le domaine de la défense, souvent même plus que le système lui-même. Ces circonstances inhabituelles nous ont amenées à modifier l'outil pour qu'il soit en mesure de les traiter.

Dans le secteur du transport, les tests ont permis de valider la capacité du XAIRS à intégrer des contraintes environnementales et légales, ainsi qu'à fournir des directives claires pour le développement de systèmes explicables. Les tests nous ont également aidés à détecter des besoins en XIA qui n'avaient jusqu'à présent pas été relevés, tels que le besoin d'avoir confiance dans l'explication et non uniquement le besoin d'avoir confiance dans la prédiction. Cette découverte a conduit à des ajustements importants dans l'outil pour mieux répondre à ces nouveaux besoins identifiés.

Dans le secteur de la santé, le XAIRS a été testé dans un cas de classification binaire où une réponse positive négative n'était pas suffisante. Le défi était de comprendre comment transformer une prédiction négative en une prédiction positive. La XIA se plaçait donc au cœur de la réussite industrielle de ce projet. Contrairement aux autres tests, cette phase n'a pas été réalisée dans le but d'améliorer l'outil, mais plutôt pour confirmer son utilité. Une analyse approfondie des retours des personnes menant le projet a été effectuée pour évaluer l'efficacité du XAIRS dans ce contexte spécifique. Ceci inclut un gestionnaire de projet et un responsable technique. Une analyse des retours est présentée à la partie 6.7.

6.4 Structure et utilité de l'outil *Explainable AI Requirements Specification*

Cette section vise à détailler la structure et les principales utilités du XAIRS, dont sa capacité à guider efficacement les ingénieurs et les parties prenantes dans le domaine complexe de la XIA. Nous explorons la structure et les applications de l'outil, démontrant comment il guide les ingénieurs et les parties prenantes dans l'identification des exigences, des contraintes et des fonctionnalités de la XIA.

Afin de faciliter la lecture, la figure 6.3 présente la table des matières du XAIRS. L'outil est aussi présenté à l'annexe I

L'outil XAIRS est structuré de manière à faciliter l'identification et la spécification des exigences en matière d'explicabilité pour les systèmes d'IA grâce à un déroulement logique et cohérent d'une analyse technique. Cette structure est conçue pour être à la fois exhaustive et intuitive. Dans cette section, nous passons au travers de cette suite logique afin de montrer son utilité pour les développeurs et les autres parties prenantes concernés.

Conformément aux standards de l'IEEE pour la spécification des exigences logicielles, l'outil intègre des sections types telles que l'introduction, la description générale, les exigences spécifiques, et les contraintes opérationnelles. Ces sections familières faciliteront l'adoption de XAIRS par les professionnels de l'informatique et de l'ingénierie.

Figure 6.3 Table des matières de l'outil *Explainable AI Requirements Specification*

Comme dans le format du SRS, la première section du XAIRS, nommée Introduction (section 1), clarifie la présente documentation, en décrivant son but, les conventions utilisées, l'audience visée, la portée du produit, à savoir du système d'IA explicable, et les références. Cette section ne comporte pas de différences avec un SRS standard, à l'exception de la description des sous-

sections qui comportent une orientation spécifique pour la XIA. Par exemple, la sous-section Portée du produit (1.4) présente la description suivante afin de guider l'utilisateur de l'outil.

« Fournir une description concise du système d'intelligence artificielle dont l'explicabilité est évaluée et de l'objectif visé. Décrire les principaux avantages, objectifs et buts de l'amélioration ou de la création de l'explicabilité au sein de ce système d'intelligence artificielle, en soulignant la manière dont ils s'alignent sur la fiabilité et la compréhension de l'utilisateur. Relier les aspects de l'IA relatifs à l'explicabilité aux objectifs organisationnels généraux ou aux stratégies commerciales. S'il existe un document distinct sur la vision et la portée détaillant le contexte plus large ou les objectifs globaux de la mise en œuvre de l'IA, il convient de s'y référer ici sans dupliquer son contenu. » (Explainable AI Requirements Specification, Annexe I)

La section *Description générale* (section 2) contient elle aussi les sous-sections que l'on retrouve dans un document SRS, soit la perspective du produit, les catégories et caractéristiques des utilisateurs, l'environnement opérationnel, les contraintes de conception et de mise en œuvre, la documentation destinée à l'utilisateur ainsi que les hypothèses et dépendances. Tout comme l'introduction, les descriptions des sous-sections possèdent une orientation spécifique pour la XIA. Pour enrichir l'établissement des exigences, réalisé à la section 4 du XAIRS, certaines de ces sous-sections ont été détaillées, notamment la sous-section portant sur les catégories et caractéristiques des utilisateurs. Celle-ci a été subdivisée en trois, suivant les directives de l'article *Merging Roles and Expertise : Redefining Stakeholders Characterization in Explainable Artificial Intelligence* (Raymond et al., 2024). Ces trois subdivisions sont (1) les catégories de parties prenantes et leur priorisation (section 2.2.1), (2) les descriptions des parties prenantes (section 2.2.2) et (3) les caractéristiques des parties prenantes (section 2.2.3). Ce découpage permet de répondre à trois besoins : une priorisation des exigences en XIA réalisée grâce à une priorisation des catégories de parties prenantes, l'identification des rôles des parties prenantes pour une clarification ultérieure de leurs besoins et finalement, une identification des connaissances des parties prenantes afin d'éventuellement cibler un niveau de vulgarisation idéal. La sous-section nommée Contraintes

de conception et de mise en œuvre a également été subdivisée en 2 sous-sections : contraintes juridiques (section 2.5) et principes éthiques (section 2.6).

La section *Interactions du système* (section 3), est unique au document XAIRS et est au cœur de la démarche d'identification des exigences et contraintes en XIA. Elle se compose de trois sous-sections, visant à déterminer les informations recherchées par les parties prenantes à différents moments du cycle de vie du système. La première, Cycle de vie du système (section 3.1), décrit l'ensemble du cycle de vie, en soulignant les étapes clés depuis sa conception jusqu'à son utilisation active par les utilisateurs finaux. Le format flexible permet aussi d'indiquer uniquement les phases jugées pertinentes en fonction de la maturité du projet. Il est toutefois pertinent de noter que, dans le domaine du génie logiciel, identifier les exigences en fonction du produit final plutôt que de se concentrer uniquement sur les phases intermédiaires est considéré comme une pratique optimale afin de diminuer les coûts et optimiser les ressources (Ruparelia, 2010). La deuxième sous-section, nommée Interactions entre les parties prenantes du système (section 3.2), décrit la manière dont les parties prenantes interagissent avec le système d'IA, pour chaque phase du cycle de vie de l'IA. Le format de tableau qui y est suggéré permet de combiner les phases du cycle de vie du système (section 3.1) et les parties prenantes (section 2.2.2) pour définir les interactions. Aussi, les contraintes juridiques (section 2.5), les principes éthiques (section 2.6) et la documentation utilisateur (section 2.7) doivent être pris en considération et cités dans la liste d'interactions possibles avec le système. Cette sous-section permet ensuite de remplir la suivante : Évaluation des besoins en information. Celle-ci identifie et catégorise les différents besoins des parties prenantes pour le système d'IA, en se concentrant sur les personnes qui ont besoin d'informations spécifiques et sur le moment auquel ces informations sont nécessaires. Cette démarche est basée sur l'hypothèse que les besoins en information proviennent des interactions entre les parties prenantes et le système d'IA.

La section qui suit, nommée *Fonctionnalités d'explicabilités* (section 4), présente un format très similaire à celui présenté dans le SRS. Pour chaque fonctionnalité identifiée, il est exigé d'y incorporer une description, un niveau de priorité et les exigences fonctionnelles. Afin de

créer des fonctionnalités, il est recommandé de suivre la démarche suivante. Premièrement, pour chaque fonctionnalité, attribuer un nom significatif et ajouter une description qui incluent les informations suivantes : les parties prenantes qui utiliseront cette fonctionnalité, l'information(s) recherchée(s), ainsi que phases du cycle de vie dans lesquelles seront intégré la fonctionnalité. Ensuite, associer le niveau de priorité en fonction de la partie prenante qui utilisera cette fonctionnalité. Cette information est définie à la section : Catégories de parties prenantes et priorités (section 2.2.1). En revanche, si cette fonctionnalité répond à une contrainte légale, peu importe le niveau de priorité de la partie prenante, il est essentiel d'indiquer un niveau de priorité critique. Les niveaux de priorités varient de basse à critique. Et finalement, il faut définir les exigences de la fonctionnalité. Pour ce faire, vous devriez utiliser les sections Environnement (section 2.3) et la phase du cycle de vie (section 3.3) afin de définir les exigences reliées à la rapidité de lecture et de compréhension des explications fournies. Aussi la caractérisation de la partie prenante (section 2.2.3) permet de définir des exigences qui dépendent des connaissances des parties prenantes, tel que les méthodes de vulgarisation ou de présentation optimales des explications.

La dernière section présente les annexes (section 5). Celle-ci contient minimalement un glossaire, tel que présenté dans les documents SRS.

6.5 Structure et utilité de l'outil Explainable AI Requirements Specification Template for Common AI Projects

Cette section vise à détailler le contenu de l'outil Explainable AI Requirements Specification Templates for Common AI Projects, ainsi que son utilité en relation avec le XAIRS. Nous décrivons la structure et les avantages des Explainable AI Requirements Specification Templates for Common AI Projects, visant à simplifier l'utilisation du XAIRS en offrant des gabarits pour surmonter sa complexité et la dépendance à l'expertise des ingénieurs.

Le XAIRS, bien que prometteur, possède quelques limitations qui pourraient rendre la communauté d'ingénierie réticente à son utilisation :

1) Complexité et surcharge d'information : l'outil offre une approche détaillée et complète, ce qui peut entraîner une complexité accrue. Cela peut occasionner des difficultés à utiliser efficacement l'outil.
2) Dépendance à l'expertise des ingénieurs : l'efficacité de l'outil dépend en partie du niveau d'expertise des ingénieurs en matière de XIA. Pour ceux qui ne possèdent pas une compréhension approfondie du domaine, il pourrait être difficile de tirer pleinement parti des utilités de l'outil.

Afin de pallier ces difficultés, nous avons mis en place quelques mesures concrètes.

En premier lieu, en plus d'une description des informations attendues pour chacune des sous-sections (tel que citer en exemple à la section 4.1 de cet article), nous avons ajouté, directement dans le XAIRS, des exemples. Ces exemples permettent aux parties prenantes de mieux comprendre l'information qui doit être communiquée dans chacune des sections et le niveau de détail approprié. Par exemple, la sous-section Portée du produit présente le cas suivant afin de guider l'utilisateur de l'outil.

« MediAI Diagnosis v3.2 est un système d'IA conçu pour faciliter les diagnostics médicaux, en particulier dans le domaine de la cardiologie. Pour que les professionnels de la santé puissent faire confiance aux suggestions de l'IA et les utiliser efficacement, il est essentiel d'améliorer sa capacité d'explication. Cela correspond à notre objectif d'intégrer des technologies avancées dans les soins de santé, en respectant à la fois la sécurité des patients et les méthodes de diagnostic innovantes. De plus amples informations sont disponibles dans le document MediAI Vision and Scope. » (Explainable AI Requirements Specification, Annexe I)

Aussi, afin d'accélérer le processus d'écriture du XAIRS et de permettre aux parties prenantes qui possèdent une expertise limitée dans le domaine de la XIA de l'utiliser efficacement, nous avons créé l'outil *Explainable AI Requirements Specification Templates for Common AI Projects*. Pour simplifier la lecture, nous nous référerons à celui-ci comme les XAIRS Templates (XAIRST). L'objectif de cet outil est de fournir des gabarits pour les sections plus longues à remplir du XAIRS et de guider les ingénieurs au travers des différents concepts et

connaissances de la XIA. Ainsi, les entreprises peuvent reprendre les différentes sections et les adapter aisément à leur projet.

Le XAIRST est présenté sous le même format que le XAIRS, suivant les bonnes pratiques définies dans le SRS. Certaines sections présentent des gabarits.

La sous-section 2.2, portant sur les classes et caractéristiques des utilisateurs, est la première à proposer des gabarits. Ces gabarits ont été repris de l'article *Merging Roles and Expertise: Redefining Stakeholders Characterization in Explainable Artificial Intelligence* (Raymond et al., 2024). Ensuite, la section *Contraintes légales* (section 2.4) fournit aussi un gabarit qui présente les contraintes légales, en termes de XIA, au Québec. Ce gabarit a d'ailleurs été écrit avec la collaboration de Maître Chassigneux qui a travaillé pendant plus de dix ans à la Commission d'accès à l'information du Québec, dont six ans à titre de juge administratif affecté à la section de surveillance. Ce gabarit ne tient pas compte de l'utilisation prévue de l'IA (exemple : système de recommandation de musique ou système de recommandation d'un traitement médical) ni du secteur d'activité (exemple : médical, bancaire, immobilier). Il est de la responsabilité de l'organisme d'ajouter des contraintes légales concernant ces aspects. Aussi, si une entreprise œuvre à l'extérieur du Québec, celle-ci devra adapter le gabarit en conséquence. Le prochain gabarit proposé (section 2.5) est celui qui présente les principes éthiques applicables aux projets XIA. Ce gabarit propose des principes et une description de ceux-ci basés sur les principes éthiques présentés par la Déclaration de Montréal sur l'Intelligence Artificielle Responsable (DMDRIA). Une entreprise pourrait choisir d'y inclure ses propres principes. Les autres sous-sections de la section 2, Description générale, ne possèdent pas de gabarits puisque celles-ci ne sont pas généralisables, mais plutôt spécifiques à chaque projet.

La section 3 du XAIRST, nommée *Interactions système*, propose un gabarit à chacune des sous-sections car c'est l'une des sections spécifiques à l'outil XAIRS, c'est-à-dire non présent dans le SRS. Le premier gabarit propose le cycle de vie d'un système d'IA, inspiré du modèle *Data-Centric AI*. Les sous-sections suivantes, soit l'interaction entre les parties prenantes du

système d'IA et l'évaluation des besoins en information, ont été élaborées en s'appuyant sur plusieurs études importantes dans le domaine de la XIA. On y retrouve notamment Langer et al. (2021), Mohseni et al. (2021), Dhanorkar et al. (2021), He et al. (2023) ainsi que Gerlings et al. (2020).

La quatrième section, intitulée *Fonctionnalités d'explicabilité*, a été élaborée en appliquant le protocole décrit à la partie 4 de cet article aux exemples fournis dans les sections antérieures. Cette section se concentre sur le développement de fonctionnalités pour un projet d'IA commun. Pour ce faire, nous avons identifié les fonctionnalités de la XIA les plus fréquentes dans la littérature. Des références telles que Haque et al. (2023), Saeed et al. (2023), Vyas (2023) ainsi que Dwivedi et al. (2023) ont été utilisées. Les fonctionnalités identifiées ne répondent pas à tous les besoins en information énumérés dans la section 3 du XAIRST. Par conséquent, ces éléments peuvent être considérés comme des besoins non satisfaits par la littérature actuelle et représentent des pistes d'exploration pour de futures recherches. Pour définir les exigences fonctionnelles des fonctionnalités, nous nous sommes appuyés sur les travaux de recherche de (Chromik et Butz, 2021), Lei et al. (2024) ainsi que Liao et al. (2020) qui ont proposé des lignes directrices pour simplifier et communiquer efficacement des explications claires aux parties prenantes.

6.6 Évaluation des outils Explainable AI Requirements Specification et Explainable AI Requirements Specification Templates for Common AI Projects

Dans le cadre de la phase finale de test des outils XAIRS et XAIRST, nous avons mené un sondage exploratoire auprès de deux professionnels en milieu industriel : un gestionnaire de projet IA et un responsable technique en développement d'IA explicable. Leur retour d'expérience a permis de recueillir des données qualitatives et quantitatives sur la pertinence, l'utilité et les limites perçues des outils. Ce retour, bien que limité en nombre de répondants, donne un aperçu préliminaire, mais éclairant de la valeur ajoutée des outils en contexte réel.

Les deux répondants ont unanimement souligné la valeur ajoutée du XAIRS, qu'ils ont chacun évalué à 4/5 pour son utilité dans la structuration d'un projet de XIA. Le XAIRS a été particulièrement bien accueilli par les équipes techniques en charge du développement de systèmes d'IA explicables. Il en est ressorti que, grâce à sa structure méthodique, le XAIRS facilite la prise en compte des exigences en matière d'explicabilité dès les premières étapes d'un projet. Selon le responsable technique interrogé, le protocole d'identification des parties prenantes a permis une couverture plus complète des profils concernés, y compris des parties souvent négligées, comme les auditeurs externes ou les utilisateurs indirects. Le gestionnaire de projet a quant à lui noté que ce processus a permis de clarifier les priorités en matière de vulgarisation, en fonction des niveaux de connaissances des parties prenantes, ce qui a mené à des décisions plus alignées avec les objectifs de clarté et de transparence.

En complément, la définition structurée des fonctionnalités de la XIA intégrée au XAIRS s'est avérée bénéfique pour guider à la fois la conception, la documentation et l'évaluation des systèmes. Les fonctionnalités identifiées ont été perçues comme des repères concrets permettant de s'assurer que les requis en XIA sont non seulement formellement définis, mais aussi vérifiables en fin de projet. À cet égard, les répondants ont mentionné que le XAIRS aide à traduire des principes souvent abstraits en pratiques tangibles, facilitant le dialogue entre les membres d'une équipe multidisciplinaire. Pour eux, cela représente un changement notable par rapport aux approches antérieures, souvent intuitives et peu formalisées. Tous deux ont estimé que cet outil avait permis de réduire le temps consacré au développement (programmation) en explicabilité de près de 40 %. En revanche, le temps consacré à remplir le XAIRS était 15% plus long que le temps habituellement consacré à l'analyse des besoins en XIA.

Par ailleurs, les répondants ont indiqué que le XAIRST joue un rôle complémentaire essentiel pour tirer pleinement profit du XAIRS. Tous deux ont décrit le XAIRST comme un élément facilitateur, voire indispensable, pour mener à bien l'analyse éthique et légale dans le cadre de projets d'IA explicable. En fournissant des gabarits actualisés et contextualisés, le XAIRST permettrait selon leurs estimations de réduire de 20 à 25 % le temps nécessaire à la production de documents de conformité. L'un des répondants a particulièrement insisté sur le fait que ces

gabarits offrent un point de départ rassurant et structuré pour des équipes n'ayant pas de ressources spécialisées dans les enjeux juridiques ou éthiques de l'IA.

Un autre avantage relevé par les participants est que le XAIRST diminue les réticences initiales à entreprendre un travail de documentation souvent perçu comme lourd et complexe. En rendant la tâche plus accessible et mieux balisée, l'outil contribue à une démocratisation des pratiques de XIA, en particulier pour les petites et moyennes organisations. Selon les propos recueillis, les gabarits jouent ici un rôle pédagogique important : ils structurent la réflexion, stimulent les bonnes questions, et favorisent une montée en compétence progressive sur les enjeux de l'explicabilité.

Pris ensemble, le XAIRS et le XAIRST ont été décrits comme apportant une plus-value notable sous plusieurs angles :

1) Structuration de pratiques concrètes : En structurant les pratiques de la XIA autour d'un cadre concret, intégrant les dimensions sociales, techniques et juridiques, ils favorisent une mise en œuvre cohérente et interdisciplinaire de l'explicabilité.
2) Pionnier en audit et normalisation de la XIA : Ces outils ouvrent la voie à une certaine standardisation de la XIA, en proposant une documentation formelle qui pourrait, à terme, servir de base à des audits ou à des certifications.
3) Harmonisation des besoins des parties prenantes : Ils assurent également une meilleure harmonisation des besoins en explicabilité entre les parties prenantes, grâce à une identification systématique de leurs profils, attentes et niveaux de compréhension.
4) Démocratisation de la XIA : Ils rendent les bonnes pratiques de la XIA accessibles à un plus large éventail d'organisations, réduisant les coûts d'entrée pour les structures ne disposant pas d'une équipe dédiée à la XIA.

Malgré ces retours très positifs, les participants ont également émis des suggestions d'amélioration, jugées nécessaires pour accroître l'impact des outils à moyen et long terme. En effet, 100 % des répondants ont exprimé le souhait de voir le XAIRS s'enrichir par l'intégration de connaissances issues de disciplines connexes, telles que la psychologie

cognitive, les sciences de l'éducation ou les neurosciences. Cette ouverture permettrait de concevoir des explications mieux calibrées aux capacités d'attention, de compréhension et de mémorisation des utilisateurs, ce qui, selon eux, pourrait améliorer de 20 à 30 % l'efficacité des communications explicatives dans certains contextes à forte complexité.

Les répondants ont aussi recommandé une personnalisation du XAIRST selon les domaines d'application, en soulignant le besoin spécifique d'un gabarit adapté au secteur de la santé. Celui-ci devrait, par exemple, inclure des parties prenantes comme les patients, les médecins, ou les comités d'éthique, et intégrer des normes juridiques spécifiques telles que les réglementations en vigueur sur la protection des données de santé. Une telle adaptation pourrait, selon leurs estimations, réduire de 30 % le temps requis pour démarrer un projet en XIA dans un domaine spécialisé, en supprimant la nécessité de personnalisation manuelle du cadre documentaire. Ce gain pourrait être considérable.

Enfin, les deux répondants ont souligné, à 100 %, l'importance d'un mécanisme de mise à jour régulière des gabarits fournis dans le XAIRST. Cette fonctionnalité est perçue comme essentielle pour maintenir la conformité des projets face à l'évolution rapide des normes, des lois et des attentes sociétales. À défaut, ils estiment que la pertinence des outils pourrait décliner significativement au bout de 18 à 24 mois, compromettant leur adoption à long terme.

6.7 Conclusion

Nous avons développé le XAIRS pour répondre aux besoins croissants en matière de la XIA. Face aux défis techniques, éthiques et légaux posés par l'intégration de l'IA dans des domaines critiques, il est impératif de disposer de méthodologies concrètes et de protocoles robustes pour assurer la transparence et la confiance dans les systèmes d'IA.

Les résultats de notre recherche montrent que le XAIRS fournit un cadre méthodologique structuré pour spécifier les exigences, les contraintes et les fonctionnalités de la XIA. En

intégrant les principes de conception de clarté, de conformité et de flexibilité, le XAIRS permet aux ingénieurs de développer des systèmes d'IA explicables qui répondent aux attentes légales et éthiques, tout en étant adaptés aux différents contextes sectoriels. L'outil facilite également l'identification et la caractérisation des parties prenantes, la définition des besoins en information, et la spécification des fonctionnalités de la XIA nécessaires.

Le XAIRS a été validé à travers plusieurs phases de tests industriels dans les secteurs de la défense, du transport et de la santé. Ces tests ont permis d'évaluer la robustesse et l'adaptabilité de l'outil, ainsi que d'apporter des améliorations itératives basées sur les retours des utilisateurs et des experts. Les résultats montrent que le XAIRS permet de structurer les pratiques de la XIA de manière concrète et multidimensionnelle, en harmonisant les besoins des parties prenantes et en facilitant l'audit et la normalisation des systèmes d'IA explicables.

En complément, le XAIRST propose des gabarits pour simplifier l'utilisation du XAIRS, en offrant un cadre accessible et détaillé pour l'analyse des enjeux éthiques et légaux. Le XAIRST réduit les réticences à entreprendre la documentation d'IA explicable en fournissant des exemples pratiques et des sections préremplies, permettant aux entreprises de gagner du temps et d'optimiser leurs ressources.

Néanmoins, des possibilités d'optimisation existent pour enrichir davantage le XAIRS et le XAIRST. Des recherches approfondies dans des domaines interdisciplinaires tels que l'éducation et la psychologie cognitive pourraient fournir des perspectives précieuses pour améliorer les méthodologies de vulgarisation et de présentation de l'information explicative. De plus, une adaptation des outils à des domaines spécifiques, tels que le secteur médical ou bancaire, permettrait de mieux répondre aux besoins particuliers de ces industries.

En conclusion, le XAIRS et le XAIRST représentent des avancées significatives dans le domaine de la XIA, offrant des cadres méthodologiques robustes et adaptables pour développer des systèmes d'IA explicables. Leur évolution continue, soutenue par des recherches

interdisciplinaires et des adaptations sectorielles, assurera leur pertinence et leur efficacité à long terme dans la promotion de l'IA de confiance.

CHAPITRE 7

DISCUSSION

7.1 Synthèse des contributions

Dans le cadre de cette recherche, nous avons exploré de manière systématique les dimensions conceptuelles, opérationnelles, réglementaires et méthodologiques de la XIA, dans un contexte québécois marqué par des attentes croissantes en matière de transparence, de responsabilité et d'éthique. Les quatre articles présentés s'inscrivent dans une démarche de recherche appliquée, cherchant à répondre aux lacunes à la fois théoriques et pratiques entourant la XIA, tout en soutenant l'appropriation de ce concept par les acteurs du milieu, notamment les ingénieurs qui développement des SIA.

Notre objectif initial était de combler le décalage persistant entre les explications générées par les techniques de la XIA et les besoins réels des parties prenantes, en vue d'améliorer la légitimité des systèmes d'IA explicables dans l'industrie québécoise. Pour faire avancer la recherche en ce sens, nous avons défini que nous souhaitions proposer une méthodologie qui permet de cerner et définir les besoins, exigences et contraintes de la XIA. Ainsi, nous avons offert des avancés multidimensionnel dans le contexte de la XIA.

Le premier article s'est intéressé à l'analyse du cadre législatif québécois concernant la XIA, mettant en lumière l'ambiguïté et le manque de précision dans les termes légaux. Nous avons proposé une interprétation concrète des obligations légales liées à la XIA, en les traduisant en recommandations pratiques adaptées au contexte de l'ingénierie et de la gestion de projets en IA. De plus, nous avons émis des recommandations en direction des entité législatives et public, afin de promouvoir la collaboration entre ses mêmes entités et les experts en IA, désireux de comprendre et de respecter les lois québécoises.

Le deuxième article a exploré l'opérationnalisation des principes éthiques en XIA, en soulignant qu'il ne suffit pas de déclarer une intention éthique (par exemple via la Déclaration

de Montréal) pour que celle-ci soit concrètement intégrée dans les systèmes. Nous y avons proposé une méthodologie, basé sur les principes normatifs, qui fournit aux ingénieurs une approche structurée pour traduire les principes éthiques en mesures concrètes applicables dans le contexte de la XIA.

Le troisième article a porté sur la caractérisation des parties prenantes dans les projets de XIA. Il a mis en évidence les limites des approches traditionnelles, soit centrées sur les rôles, soit sur les expertises, en montrant que cette dichotomie ne suffit pas à saisir la complexité des besoins en information. Il a proposé une approche fusionnée, combinant rôle et expertise, pour une identification plus fine des exigences de la XIA. En relation avec les objectifs de cette recherche, nous y avons introduit un protocole qui identifie et catégorise les différentes parties prenantes en relation avec un système de XIA. Nous avons aussi proposé une méthodologie structurée pour évaluer les connaissances applicables au domaine de l'IA des parties prenantes. Finalement, nous avons apporté une approche structurée pour hiérarchiser les besoins en XIA, basé sur une priorisation des parties prenantes, comblant ainsi une lacune dans les démarches actuelles.

Le quatrième article constitue l'aboutissement de cette recherche. Il propose une méthodologie complète pour la spécification des requis en XIA, désignée sous le nom de XAIRS (*Explainable Artificial Intelligence Requirements Specification*). Ce dernier ne peut être compris sans les apports fondamentaux des trois articles précédents, qui ont permis d'établir les fondations conceptuelles, normatives et organisationnelles nécessaires.

En effet, pour concevoir le XAIRS, il a d'abord fallu :

- Clarifier le cadre juridique de la XIA (article 1), afin d'identifier les obligations explicites et implicites pouvant être traduites en exigences fonctionnelles ;
- Traduire les engagements éthiques en indicateurs opérationnels (article 2), ce qui permet d'ancrer les principes dans des pratiques techniques concrètes ;
- Identifier, caractériser et prioriser les parties prenantes (article 3), en tenant compte à la fois de leur rôle dans le cycle de vie des SIA et de leur niveau d'expertise en IA.

C’est sur la base de ces trois piliers que le quatrième article a pu développer un outil structurant, permettant aux équipes d’ingénierie de spécifier les exigences, contraintes et besoins de la XIA de manière rigoureuse, contextualisée et documentée. Le XAIRS agit ainsi comme un point de convergence des contributions antérieures : il rend opérationnelles les recommandations juridiques, éthiques et méthodologiques précédemment formulées, tout en offrant un canevas pour guider concrètement le travail de conception dans les projets d’IA.

7.2 Analyse transversale des résultats

Chacun des articles s’inscrit dans une dynamique qui vise à outiller les ingénieurs et professionnels de l’IA pour traduire les exigences abstraites de la XIA en éléments concrets, contextualisés et opérationnalisables. Cette démarche repose sur quatre principes transversaux qui reviennent de manière constante dans l’ensemble de la recherche : la traduction des exigences abstraites en artefacts d’ingénierie, l’identification comme geste fondateur, la contextualisation comme condition de pertinence, et enfin le rôle structurant de la méthodologie comme garant de la légitimité.

Premièrement, l’un des apports majeurs de cette recherche réside dans la capacité à traduire des exigences juridiques, éthiques et sociales souvent formulées de façon générale, voire abstraite, en exigences techniques directement mobilisables par les équipes d’ingénierie. Le premier article (voir chapitre 3) en constitue une première démonstration, en interprétant les zones grises de la Loi 25 pour en extraire des recommandations applicables à la conception de systèmes IA. Le deuxième article (voir chapitre 4) poursuit cette logique sur le plan éthique, en montrant comment des principes tels que la transparence peut être déclinés en indicateurs concrets et suivis dans un projet. Finalement, dans l’article 4, les exigences sont intégrées dans un cadre formel de spécification, le XAIRS, permettant leur mise en œuvre cohérente dès les premières étapes du développement. Ainsi, la XIA n’est pas présentée comme un idéal à atteindre, mais comme un processus d’ingénierie structuré, ancré dans les réalités techniques.

Deuxièmement, un motif méthodologique central émerge : l'identification comme préalable incontournable. Avant de concevoir, d'expliquer ou d'évaluer quoi que ce soit, il est nécessaire d'identifier précisément ce qui est attendu, de qui, dans quel contexte et avec quelles contraintes. Cette logique se retrouve dans l'identification des obligations légales dans l'article 1, des principes éthiques dans l'article 2, et des parties prenantes dans l'article 3. Chaque article repose sur cette idée fondatrice selon laquelle toute démarche d'explicabilité doit commencer par une phase rigoureuse de repérage : des normes, des principes, des acteurs et de leurs besoins. L'article 4, en systématisant cette étape dans sa méthodologie XAIRS, confirme que l'identification n'est pas une étape périphérique mais bien une condition de validité de toute approche XIA.

Troisièmement, les quatre articles partagent une même conviction : l'explicabilité n'a de sens que si elle est contextualisée. Une même explication peut être pertinente ou inutile selon le profil du destinataire, son rôle dans le système, son niveau d'expertise, mais aussi selon le cadre légal ou les valeurs sociales en jeu. L'article 1 met en lumière les spécificités du contexte québécois en matière de lois sur la protection des renseignements personnels, tandis que l'article 2 insiste sur l'ancrage des principes dans des réalités contextuelles (par exemple, sectorielle tel que la santé, finance, justice). L'article 3 propose une méthode de caractérisation des parties prenantes qui rend possible cette contextualisation en croisant rôle et connaissance. Enfin, l'article 4 intègre systématiquement cette exigence dans son protocole, en liant chaque exigence à un contexte d'application clairement défini. La contextualisation apparaît ainsi comme une réponse aux faiblesses des approches génériques en XIA, et comme une garantie de la pertinence.

Enfin, cette recherche illustre que la XIA, pour être légitime et durable, doit s'inscrire dans une méthodologie rigoureuse. Trop souvent perçue comme une notion floue ou comme un objectif symbolique, la XIA ne peut gagner en crédibilité qu'à travers des démarches reproductibles, transférables et évaluables. C'est ce que démontre l'article 2 en proposant une méthode d'opérationnalisation éthique, l'article 3 en introduisant une procédure structurée de caractérisation des acteurs, et surtout l'article 4 en formalisant une méthodologie complète de

spécification. Le XAIRS devient ainsi le point de convergence des travaux antérieurs, réunissant les acquis juridiques, éthiques et méthodologiques en un seul outil. Il traduit une posture fondamentale : la XIA ne peut pas se contenter de principes, elle doit reposer sur des pratiques. Et ces pratiques ne peuvent exister sans méthodologie.

En somme, cette analyse montre que les quatre articles forment un tout cohérent, structuré par une logique de progression. Ensemble, ils construisent une réponse à la question centrale de la recherche : comment combler le décalage persistant entre les explications générées par les techniques de la XIA et les besoins réels. En traduisant l'abstrait, en identifiant rigoureusement les attentes, en contextualisant les explications et en structurant les démarches, cette recherche contribue à bâtir une ingénierie de la XIA de confiance.

7.3 Limites de la recherche

Cette recherche a permis de proposer une méthodologie structurée et multidimensionnelle pour cerner, formaliser et prioriser les exigences en matière d'intelligence artificielle explicable (XIA). Néanmoins, elle comporte plusieurs limites qu'il convient de reconnaître, tant sur le plan méthodologique, empirique que conceptuel.

Premièrement, certaines contributions, notamment celles formulées dans les articles juridiques et éthiques, n'ont pas pu faire l'objet d'une évaluation indépendante dans un contexte industriel. Ces recommandations ont été intégrées et testées indirectement dans le cadre du quatrième article, à travers la validation globale de la méthodologie XAIRS. Toutefois, en raison de leur nature ancrée dans l'interprétation normative et l'engagement éthique, il n'a pas été possible de mener une évaluation isolée ou comparative (par exemple par des tests A/B), ce qui limite la mesure directe de leur impact individuel. Leur efficacité repose donc sur une validation d'ensemble, dans le cadre d'un processus intégré.

Deuxièmement, bien que la méthodologie proposée soit conçue pour être générique et adaptable, la recherche n'a pas été ancrée dans un secteur d'activité spécifique. Or, les

exigences d'explicabilité varient considérablement selon les domaines sectoriels (exemples : santé, finance, transport, justice) en fonction des cadres réglementaires, des pratiques organisationnelles et des attentes des utilisateurs finaux. La méthodologie permet d'identifier ces exigences spécifiques, mais elle ne fournit pas de catalogue de requis sectoriels prédéfinis. Cette absence volontaire de spécialisation sectorielle constitue une limite si l'on cherche des solutions prêtes à l'emploi pour un domaine donné.

Troisièmement, la recherche a été menée dans un contexte normatif et culturel particulier, celui du Québec, marqué par des attentes spécifiques en matière de transparence, de responsabilité et de respect de la vie privée. Les recommandations issues de l'analyse juridique, en particulier celles liées à la Loi 25, sont donc partiellement dépendantes de ce contexte. Leur transposition dans d'autres juridictions nécessitera une adaptation fine aux cadres législatifs et sociopolitiques locaux.

Enfin, comme toute démarche visant à traduire des principes abstraits en artefacts d'ingénierie, cette recherche s'est centrée sur ce qui peut être structuré, spécifié, documenté. Si cette posture permet d'opérationnaliser la XIA, elle laisse en partie de côté des dimensions plus subjectives, telles que la réception des explications par les utilisateurs, leur confiance réelle ou leurs émotions face à l'IA. Une exploration plus fine de ces aspects relèverait de démarches complémentaires, issues par exemple des sciences sociales ou de l'éthique appliquée.

Ces limites ne remettent pas en cause la validité des résultats obtenus, mais elles encadrent leur portée et leur champ d'application. Elles dessinent aussi des pistes pour des travaux futurs visant à étendre, adapter ou compléter les apports de cette recherche.

7.4 Pistes de recherche futures

La présente recherche ouvre plusieurs pistes prometteuses pour poursuivre l'exploration et l'approfondissement des enjeux liés à l'intelligence artificielle explicable (XIA), notamment dans des contextes d'ingénierie appliquée.

Une première piste concerne le développement de méthodes d'évaluation standardisées pour les explications générées. Aujourd'hui, même lorsque des exigences sont correctement spécifiées, il demeure difficile d'évaluer objectivement la qualité ou la pertinence des explications offertes par les systèmes XIA. Une suite logique à cette recherche serait donc de concevoir des outils ou des métriques d'évaluation ancrées dans les exigences formulées via XAIRS, capables de prendre en compte à la fois le rôle, le niveau de connaissance et les intentions des parties prenantes. Cela permettrait de renforcer la valeur pragmatique de la XIA dans des contextes professionnels concrets.

Une deuxième piste prometteuse consisterait à déployer et tester la méthodologie XAIRS dans des secteurs spécifiques, par exemple en santé, en finance ou en droit. Chaque domaine comporte ses propres contraintes réglementaires, enjeux éthiques et attentes en matière d'explicabilité. Si la présente recherche propose une méthodologie générique et flexible pour identifier ces exigences, des travaux futurs pourraient explorer, documenter et comparer la manière dont les exigences XIA se formulent, se priorisent et se traduisent différemment selon les contextes sectoriels. Cela contribuerait à constituer une base de cas d'usage concrets et à guider les ingénieurs dans des environnements complexes.

Enfin, une troisième piste consisterait à étudier plus en profondeur le lien entre les exigences XIA et les techniques explicatives elles-mêmes. La méthodologie actuelle reste volontairement agnostique sur les outils XIA employés, afin de préserver sa généricité. Or, il serait pertinent d'examiner si certains types d'exigences (par exemple, juridiques, pédagogiques ou interactives) appellent l'utilisation de certaines familles d'approches (explications post-hoc locales, visualisations, contre-exemples, etc.). Une telle recherche permettrait de créer des ponts plus solides entre la phase de spécification des besoins et les choix technologiques ultérieurs, facilitant ainsi le travail d'implémentation.

Ces pistes de recherche futures ne visent pas à corriger les limites de l'approche proposée, mais à l'inscrire dans un cycle itératif d'enrichissement, où la méthodologie XAIRS servirait à la fois de cadre initial, de base d'apprentissage, et de point d'ancrage.

CONCLUSION

Au Québec, le déploiement de l'IA soulève des enjeux croissants de transparence, d'éthique et de responsabilité. Malgré l'entrée en vigueur de la Loi 25 et la multiplication des cadres de gouvernance, la XIA demeure difficile à opérationnaliser dans les milieux industriels. Les approches existantes se concentrent majoritairement sur les aspects techniques des algorithmes, sans répondre pleinement aux besoins concrets des parties prenantes. Ce constat met en évidence un écart persistant entre les explications produites par les techniques de XIA et les attentes réelles des acteurs impliqués. C'est dans ce contexte que cette recherche s'inscrit : elle vise à combler ce décalage en proposant une méthodologie intégrée pour cerner, formaliser et prioriser les exigences, contraintes et besoins en XIA, adaptée au contexte québécois et alignée sur les principes de l'IA de confiance.

Cette recherche a proposé une approche structurée et multidimensionnelle pour opérationnaliser la XIA dans le contexte québécois. À travers quatre articles complémentaires, elle a permis d'explorer les dimensions légales, éthiques, organisationnelles et méthodologiques nécessaires à la conception d'une IA de confiance. L'ensemble de ces travaux converge vers la création d'un cadre cohérent, ancré dans les réalités normatives et industrielles du Québec.

Le premier article a clarifié le cadre juridique applicable à la XIA, en traduisant les obligations de la Loi 25 en recommandations concrètes pour les ingénieurs et décideurs. Le second a proposé une méthode d'opérationnalisation des principes éthiques, permettant de passer de l'intention à la pratique. Le troisième a introduit un protocole pour identifier, caractériser et prioriser les parties prenantes, rendant possible une compréhension fine et contextualisée de leurs besoins. Enfin, le quatrième a consolidé ces acquis dans la méthodologie XAIRS (Explainable Artificial Intelligence Requirements Specification), un outil complet de spécification des requis en XIA. En réunissant ces contributions, la recherche comble un vide important entre les principes abstraits et leur mise en œuvre technique. Elle démontre que la

XIA ne relève pas seulement d'un objectif éthique ou réglementaire, mais d'un véritable processus d'ingénierie de la confiance, reproductible et évaluable.

Les retombées de cette recherche se situent autant sur le plan scientifique que pratique. Sur le plan académique, la méthodologie XAIRS constitue une contribution originale à l'ingénierie de la XIA, en reliant les perspectives légales, éthiques et techniques à travers un cadre unifié. Sur le plan appliqué, elle fournit aux ingénieurs, gestionnaires et responsables de conformité un outil concret pour traduire les obligations de transparence en artefacts d'ingénierie. En intégrant la XIA au cœur du cycle de vie des systèmes, cette approche encourage une collaboration interdisciplinaire entre ingénieurs, juristes et décideurs publics, et positionne le Québec comme un territoire d'innovation responsable.

Bien que cette recherche ait permis de structurer une méthodologie complète pour la spécification des exigences en XIA, certaines limites encadrent la portée de ses résultats. La validation empirique s'est déroulée dans un contexte industriel unique, et l'ancrage dans le cadre québécois, bien que pertinent, en restreint la généralisation internationale. Par ailleurs, la recherche s'est centrée sur les dimensions formalisables de la XIA, laissant en partie de côté la perception humaine et émotionnelle des explications produites. Ces limites invitent à prolonger les travaux vers une exploration plus sectorielle, plus comparative et plus centrée sur l'expérience utilisateur.

En conclusion, cette recherche soutient une vision de la XIA comme discipline d'ingénierie ancrée dans la société, au croisement du droit, de l'éthique et de la technique. Elle contribue à transformer l'explicabilité, souvent perçue comme un idéal abstrait, en une pratique concrète et mesurable, au service de la confiance publique. L'enjeu n'est plus seulement de rendre les systèmes intelligents explicables, mais de rendre les explications socialement intelligibles. C'est-à-dire compréhensibles, pertinentes et légitimes pour ceux qui en dépendent. Dans cette perspective, XAIRS représente une avancée vers une IA québécoise de confiance et humaine.

ANNEXE I

EXPLAINABLE AI REQUIREMENTS SPECIFICATION

Cette annexe introduit le document intitulé « Explainable AI Requirements Specification ». Il s'agit de l'un des livrables de cette thèse qui a été présenté au chapitre 6. Le document est présenté en suivant les normes d'édition des documents de l'IEEE.

Explainable AI Requirements Specification for <Project>

Version 0.1 approved

Prepared by <authors>

<organization>

<date created>

Table of Contents

Revision History

Name	Date	Reason For Changes	Version
Camélia Raymond	*2023/12/06*	• *Reviewed section 2.2.1* • *Modified section 2.2.2 by adding a new stakeholder* • *Created section 2.2.3*	*0.1*

Acronyms

Abbreviation	Definition
AI	*Artificial Intelligence*
XAI	*Explainable Artificial Intelligence*

1. Introduction

1.1 Purpose

< Identify the artificial intelligence system whose needs, requirements, and constraints in terms of explainability are assessed in this document, including its revision or release number. Describe the scope of the AI application covered by this assessment, particularly if this evaluation addresses only a part of the system or a specific subsystem. If a document already contains its information, it is also possible to refer to it. >

Example : This document assesses the explainability needs, requirements, and constraints of the AI system MediAI Diagnosis v3.2. The scope of this evaluation covers the AI's application in providing diagnostic suggestions based on patient data, specifically its cardiology module. The goal is to enhance the system's explainability, ensuring healthcare professionals can understand and trust the AI's diagnostic processes.

1.2 Document Convention

< Describe any standards or typographical conventions that were followed when writing this document, such as fonts or highlighting that have special significance. >

1.3 Intended Audience and Reading Suggestions

< Describe the different types of reader that the document is intended for, such as developers, project managers, marketing staff, users, testers, and documentation writers. Describe what the rest of this SRS contains and how it is organized. Suggest a sequence for reading the document, beginning with the overview sections and proceeding through the sections that are most pertinent to each reader type. >

Example : This document is intended for software developers, project managers, healthcare professionals, quality assurance testers, and documentation writers. It begins with an overview of the AI system, followed by detailed explainability requirements. Readers should start with the sections most relevant to their role: developers and testers with technical specifications, and healthcare professionals with explainability aspects.

1.4 Product Scope

< Provide a concise description of the artificial intelligence system being evaluated for explainability and its intended purpose. Outline the key benefits, objectives, and goals of enhancing explainability within this AI system, emphasizing how these align with trustworthiness and user comprehension. Connect the explicability aspects of the AI to overarching organizational objectives or business strategies. If there is a separate vision and scope document detailing the broader context or the comprehensive goals of the AI implementation, refer to it here without duplicating its contents. >

Example : "MediAI Diagnosis v3.2" is an AI system designed to assist in medical diagnostics, with a focus on cardiology. Enhancing its explainability is crucial for healthcare professionals to trust and effectively use the AI's suggestions. This aligns with our goal to integrate advanced technology into healthcare, adhering to both patient safety and innovative diagnosis methods. Further details are available in the "MediAI Vision and Scope" document.

1.5 References

< List all documents, papers, or web resources that this assessment document references. These may include academic research papers on AI explainability, standards for trustworthy AI, guidelines for ethical AI design, specific industry or technical standards relevant to AI, documentation of AI systems being evaluated, or any vision and scope documents that provide context to this assessment. Ensure to provide comprehensive details for each reference, including title, author, version number, date, and source or location, to enable the reader to access a copy of each. >

RAYMOND, Camélia, RATTÉ, Sylvie, et DAOUST, Marc-Kevin. Merging roles and expertise: Redefining stakeholder characterization in explainable artificial intelligence. In : 2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON). IEEE, 2024. p. 1-7.

RAYMOND, Camélia, DAOUST, Marc-Kevin, et RATTÉ, Sylvie. Le développement d'une IA explicable: entre principes éthiques généraux et mesures concrètes. Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie, 2025, vol. 64, no 1, p. 81-99.

2. Overall Description

2.1 Product Perspective

< Describe the context and background of the artificial intelligence system whose explainability is being evaluated in this document. Indicate whether this AI system is an advancement within an existing family of AI products, a replacement for previous systems with enhanced explainability features, or a new, standalone product focusing on explicability in AI. If this assessment is for a component of a larger system, illustrate how the explainability requirements of this AI component align with the broader system's functionality, and identify the interfaces between the two. Including a simple diagram that highlights the major components of the overall AI system, the interconnections of subsystems focused on explicability, and external interfaces can be very useful. >

Example : "MediAI Diagnosis v3.2" represents an advanced iteration within the MediAI product family, specifically designed to enhance explainability in medical diagnostics. This AI system is not just an upgrade but a significant advancement over previous versions, integrating more sophisticated explainability features. It functions as a key component of a larger healthcare information system, seamlessly interfacing with patient management software. The AI's explainability features are aligned with the broader system's objective of providing transparent and trustworthy medical diagnostics. A diagram is included to illustrate the AI's integration within the larger system and its interface points.

2.2 User Classes and Characteristics

< Identify the various user classes that you anticipate will use this product and will interact with the product. User classes may be differentiated based on multiple factors. Describe the pertinent characteristics of each user class. Distinguish the most important user classes for this product from those who are less important to satisfy. >

< In this section, we recommend employing the framework Roles and Knowledge in Explainable AI (XAI). This Framework categorizes user classes based on their roles in relation to the system, enabling the prioritization of users and their characterization according to their relevant knowledge in XAI. We demonstrate the use of this framework by creating the content of the following subsections: Stakeholders Priority, User Classes and User Characteristics.

While this framework provides a solid starting point, we advise adapting and detailing it to the level deemed necessary for your specific project. This customization will ensure that the framework aligns precisely with the unique requirements and context of your initiative.

For more detailed information on this framework and guidance on its adaptation, please refer to the following article: [ajouter lien vers mon article]. >

2.2.1 Stakeholder Class and Priority

< Following the Roles and Knowledge in Explainable AI (XAI), this subsection categorizes the various stakeholders (Stakeholder Classes) expected to interact with the AI system. Each class is assigned a priority level, reflecting the significance of their needs and their impact on the system's design and features. This will allow a larger prioritization of the stakeholders that will be useful to prioritize the explainability features (section 4). >

Table 2.2 : Stakeholder classes and priority.

#SC	Stakeholder Class	Description	Priority
SC1			
SC2			

Example : See Table 2.1 in Templates AI Explainability Requirements Specification for Common AI Projects.

2.2.2 Stakeholders Descriptions

< Provide descriptions of each stakeholder within the defined classes. The descriptions include their specific roles with the AI system. If you have the knowledge of the specific resources that will be taking on those roles, they could also be included in the description. This detailed profiling assists in understanding the diverse perspectives and needs within the user base, facilitating a more user-centric development approach. >

Table 2.3 : Stakeholders descriptions.

#S	Stakeholder	Description
SC1 - Stakeholder class 1		
S1		
S2		

Example : See Table 2.2 in Templates AI Explainability Requirements Specification for Common AI Projects.

2.2.3 Stakeholders Characteristics

< Following the framework Roles and Knowledge in Explainable AI (XAI), we assess stakeholders based on their knowledge levels in key areas relevant to the AI system. These areas include understanding of AI technology, data management, and the environmental

contexts in which the system will operate (see section 5.2 for the descriptions of the knowledge types and degree of knowledge). By mapping out the knowledge profiles of different stakeholders, we can tailor the system's features and interfaces to meet the varied expertise levels, ensuring accessibility and effectiveness of the explanations. >

Table 2.3 : Degrees of knowledge for each stakeholder, by knowledge type.

		K1 - AI		K2 - Data		K3 - Env.	
#S		K1.a	K1.b	K2.a	K2.b	K3.a	K3.b
Stakeholder class 1							
S1							
S2							

Example : See Table 2.3 in Templates AI Explainability Requirements Specification for Common AI Projects.

2.3 Operating Environment

< Describe the combined technical and physical environments where the AI system will be utilized, focusing on explainability. Include hardware, operating systems, and any essential software integrations. Also, highlight the physical context, such as in an emergency room in a hospital, where the system's explainability must be swift and clear to meet urgent needs. Both environments play a critical role in shaping the AI's explainability features for effective use in specific settings. >

Example : MediAI Diagnosis v3.2" is designed to operate within both complex technical and demanding physical environments. Technically, it runs on high-performance servers equipped with the latest processors and substantial memory capacity, ensuring rapid data processing. The system is compatible with multiple operating systems, including Windows and Linux, and integrates seamlessly with existing hospital information systems and electronic health records. Physically, the AI system is primarily deployed in high-pressure areas like hospital emergency rooms. The system will also be available in a non-high-pressure area so that the medical team can learn how to use it, can test it and gain confidence in it.

2.4 Legal Constraints

< This section addresses the legal constraints related to the explainability requirements of the AI system. It explores the legal mandates and regulations that necessitate transparency and understandability in AI decision-making processes. The section will cover laws and

guidelines at various levels – local, national, and international – that pertain to the rights of individuals and organizations to understand how AI-based decisions are made. >

Table 2.4 : Transparency legal constraints applicable to the project

#C	Region	Law	Art.	Description
C1				
C2				

Example : This section delineates the legal constraints specific to Quebec that are pertinent to the explainability and transparency requirements of the MediAI Diagnosis v3.2 system. These constraints stem from laws and regulations at the provincial and national levels that govern AI systems, particularly in healthcare.

See Table 2.4 in Templates AI Explainability Requirements Specification for Common AI Projects.

2.5 Ethical Principles

< This section delves into the ethical principles and considerations relevant to the AI system. It discusses the moral principles and values that should guide the design, development, and deployment of the AI, such as fairness, transparency, non-discrimination, and respect for user autonomy. If these ethical considerations have been previously studied and documented by the company in another document, it is advisable to directly refer to that document for a comprehensive understanding. >

Table 2.5 : Ethical principles applicable to the project

#P	Principle	Description regarding the implementation
P1		
P2		

Example : This section addresses the ethical constraints and considerations that are integral to the development and application of the "MediAI Diagnosis v3.2" system. Given the critical nature of medical diagnostics, it is essential to align the system with core ethical principles.

See Table 2.5 in Templates AI Explainability Requirements Specification for Common AI Projects.

2.6 User Documentation

< Define the components of user documentation (such as user manuals, online help, and tutorials) that will be delivered with the AI software. These components should focus on elucidating the AI's decision-making processes, ensuring that users can understand and effectively interact with the system. Identify the documentation delivery formats and standards, with a focus on specific requirements related to explainability and trust in AI. >

Example : This section outlines the user documentation components that will accompany the MediAI Diagnosis v3.2 software. These materials are designed to enhance user understanding and effective interaction with the AI system, focusing on explainability and trust.

- *User Manuals: Comprehensive manuals detailing system functionality, user interface navigation, and explanation of AI decision-making processes.*
- *Online Help System: An integrated online help feature within the software, providing immediate assistance and information on demand.*
- *Interactive Tutorials: Engaging and interactive tutorials guiding users through the system's features and AI decision-making process.*
- *Documentation Formats and Standards: The documentation will be available in various formats to cater to different user preferences and accessibility needs.*

2.7 Assumptions and Dependencies

< List any assumptions (as opposed to known facts) that could affect the requirements stated in the SRS for an explainable AI system. These might include dependencies on third-party or commercial components specific to AI explainability, issues related to the development or operating environment, or constraints relevant to transparency and trust in AI. The success of the project may be impacted if these assumptions are incorrect, unshared, or subject to change. Also, identify any dependencies the project has on external factors, such as AI components intended for reuse from another project, focusing on their impact on the explainability and trustworthiness of the AI, unless already documented elsewhere (for example, in a vision and scope document or the project plan). >

Table 2.6 : Critical assumptions underlying the project.

#A	***Assumption***	***Description***
A1		
A2		

Table 2.7 : Critical dependency of the project.

#D	***Dependency***	***Description***
D1		
D2		

Example : This section outlines the assumptions and dependencies that could influence the requirements and successful implementation of the MediAI Diagnosis v3.2 system, particularly in regard to its explainability and trustworthiness. The success of the MediAI Diagnosis v3.2 project is contingent upon these assumptions holding true and dependencies being met. Significant changes in any of these areas could affect the system's development, functionality, and acceptance among users.

Table 2.6 : Critical assumptions underlying the project.

#A	Assumption	Description
A1	Data quality and availability	We assume continuous access to high-quality, diverse patient data, which is crucial for the accuracy and reliability of the AI's diagnostic recommendations.
A2	Technology advancements	The system's design is based on the assumption of steady advancements in AI and machine learning technologies, which could enhance its diagnostic capabilities and explainability over time.
A3	User technical proficiency	We assume a basic level of technical proficiency among the medical professionals using the system, impacting how they interact with and understand the AI's functionalities.
A4	Regulatory compliance	The assumption that current regulatory standards regarding AI in healthcare will remain stable during the development and initial deployment phases.

Table 2.7 : Critical dependency of the project.

#D	Dependency	Description
D1	Third-party XAI tools	The system's ability to provide clear and comprehensive explanations is partly dependent on third-party tools or components designed for AI explainability.
D2	Healthcare IT Infrastructure	The performance and integration of the AI system are dependent on the existing IT infrastructure in healthcare settings, including hardware, software, and network capabilities.

3. System Interactions

3.1 System Life Cycle

< This section of the document outlines the life cycle of the AI system, highlighting key stages from its conception to its active use by end-users. We encourage you to represent this life cycle with a diagram for better visual comprehension and, if needed, to provide detailed descriptions of each phase to ensure thorough understanding and effective implementation of explainability features throughout the AI system's life cycle. >

Example : See Graph 3.1 in Templates AI Explainability Requirements Specification for Common AI Projects. It is recommended that you provide your own graph. You can use the one provided and modify it so it represents your system.

3.2 System Stakeholders Interactions

< This section describes how stakeholders interact with the AI system, for each phase of the AI lifecycle, emphasizing the clarity and understandability of these interactions. It covers the different user roles and their specific ways of engaging with the AI, focusing on the presentation of AI decisions and feedback mechanisms. >

Table 3.1 : Stakeholders interactions with the AI system.

#	Stakeholder	Interaction(s)
Lifecycle phase 1		
I1		
I2		

Example: See Table 3.1 in Templates AI Explainability Requirements Specification for Common AI Projects. You should take into consideration the Legal Constraints (section 2.5), the Ethical Principles (section 2.6) and the User Documentation (section 2.7) and add the necessary interactions. It's also recommended to add the ID of those considerations so we can ensure that they are all represented in the interactions.

For instance, to take into consideration P1. Fairness and non-discrimination, we would add the following interaction:

User Class : Public

Interaction : Look at a report underling the fairness and non-discrimination results of the system. (P1)

3.3 Information Needs Assessment

< This section identifies and categorizes the different stakeholder requirements for the AI system, focusing on who needs specific information and when these are required. This section also introduces the Explainability Feature (refer to section 4), which enables designated stakeholders to access specific information at a given phase of the project. Should a cell in this column be left blank, it indicates that Explainable AI (XAI) is not the source for providing this information in this project. >

Table 3.2 : Stakeholder information requirements during AI system lifecycle, and what Explainability Feature (section 4) is available to provide this information.

#Inf	**Stakeholder**	**What (information needed)**	**When (lifecycle phase)**	**XAI Feat.**
Inf1				4.1
Inf2				

Example : See Table 3.2 in Templates AI Explainability Requirements Specification for Common AI Projects. If there have been additions or deletions to interactions from the common template (Table 3.1), ensure to accordingly modify the template for this section.

4. Explainability Features

< This section organizes the functional requirements of the product by explainability system features, which are the key services provided by the product in terms of explainability. This section can be organized by use case, operation mode, user class, object class, functional hierarchy, or combinations thereof, whatever makes the most logical sense for your product in the context of explainability. This section should reflect the explainability features that will allow every stakeholder to get the information they need, as presented in the section 3.3. >

4.1 Explainability Feature 1

< Do not use "System Feature 1." State the feature name in just a few words. >

Example : 4.1 Provide AI Prediction Signification Description to the Affected Party

4.1.1 Description and Priority

<Provide a short description of the feature and indicate whether it is of Very High, High, Medium, or Low priority, based on the 2.3.2 Stakeholders Priority. If a feature meets a legal requirement, it should be assigned a Very High priority, regardless of the Stakeholders Priority outlined in section 2.3.2.>

Example : This feature is dedicated to explaining the significance and implications of AI predictions to the affected party. It is designed to demystify the outcomes provided by the AI, explaining what the predictions mean in a practical, real-world context.
Priority : High (see section 2.3.1 for more details)

4.1.2 Functional Requirements

< Detail the specific functional requirements associated with this explainability feature. These could be : the moment (lifecycle phase) where the explanation is needed (identify in 3.3), the characteristics of the stakeholder that should be taken into consideration for the design of the explanation (identify in 2.3.3), design and implementation constraints, if applicable for this use case (identify in 2.3.3). Requirements should be concise, complete, unambiguous, verifiable, necessary, and directly related to explainability. >

< Each requirement should be uniquely identified with a sequence number or a meaningful tag of some kind. >

Example :
REQ1 - Availability : The explanation must be available at Use the AI system phase (see section 3.3 for more details).
REQ2 - Environment :

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.
REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.
REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.
REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.
REQ4 - Simplified domain explanations for non-expert : The system must provide explanations that are easily understandable by a party who lacks knowledge in domain expertise.
REQ4.1 - Explanations must avoid domain expertise jargon and instead use clear, simple language that is accessible to non-domain stakeholders.
REQ4.2 - Incorporate a glossary or help section for terms and concepts related to the expertise domain, accessible directly within the explanation interface.
REQ4.3 - Where appropriate, use domain-expertise analogies or comparisons to familiar concepts to aid understanding.
REQ5 - Presentation format for AI non-technical and non-domain expert stakeholders : [ajouter une description]
REQ5.1 - Provide AI and domain-expertise explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.
REQ5.2 - Use clear and simple graphics to represent AI and domain-expertise concepts. These graphics should be accompanied by captions and explanations in simple language.
REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the domain-expertise and AI explanations.
REQ6 - Prediction limits : Present not only the significance of the prediction, but also what it's not, to help the stakeholders understand the limitations of the system.

4.2 Explainability Feature 2 (and so on)

Example : For complete examples, see section 4 in Templates AI Explainability Requirements Specification for Common AI Projects.

Appendix A: Glossary

< Define all the terms necessary to properly interpret the AIXRS, including acronyms and abbreviations. You may wish to build a separate glossary that spans multiple projects or the entire organization, and just include terms specific to a single project in each AIXRS. >

Example : See Appendix A: Glossary in Templates AI Explainability Requirements Specification for Common AI Projects.

ANNEXE II

EXPLAINABLE AI REQUIREMENTS SPECIFICATION FOR COMMON AI PROJECTS

Cette annexe introduit le document intitulé « Explainable AI Requirements Specification for Common AI Projects ». Il s'agit de l'un des livrables de cette thèse qui a été présenté au chapitre 6. Le document est présenté en suivant les normes d'édition des documents de l'IEEE.

Explainable AI Requirements Specification Template for Common AI Projects

Version 1.0 approved

Prepared by

Camélia Raymond, Master student at École de Technologie Supérieur

Sylvie Ratté, Professeure at École de Technologie Supérieur

Marc-Kevin Daoust, Professeur enseignant at École de Technologie Supérieur

École de Technologie Supérieure

Created on 2023-12-06

Table of Contents

Revision History of the Document

Name	Date	Reason For Changes	Version
Camélia Raymond	2025-01-10	Document publication	1.0

Revision History of the Tables

Table Num	Last Modification Date	Last Changes	Version
2.1	2023/12/06	Created the template	1.0
2.2	2023/12/06	Created the template	1.0
2.3	2023/12/06	Created the template	1.0
2.4	2023/12/06	Created the template	1.0
3.1	2023/12/06	Created the template	1.0
3.2	2023/12/06	Created the template	1.0

Acronyms

Abbreviation	Definition
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
ETS	École de Technologie Supérieure
DCAI	Data-Centric AI

1. Introduction

1.1 Purpose

This section is intentionally left blank for project-specific content. To be completed by the project team.

1.2 Document Convention

This section is intentionally left blank for project-specific content. To be completed by the project team.

1.3 Intended Audience and Reading Suggestions

This section is intentionally left blank for project-specific content. To be completed by the project team.

1.4 Product Scope

This section is intentionally left blank for project-specific content. To be completed by the project team.

1.5 References

1. RAYMOND, Camélia, RATTÉ, Sylvie, et DAOUST, Marc-Kevin. Merging roles and expertise: Redefining stakeholder characterization in explainable artificial intelligence. In : 2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON). IEEE, 2024. p. 1-7.
2. RAYMOND, Camélia, DAOUST, Marc-Kevin, et RATTÉ, Sylvie. Le développement d'une IA explicable: entre principes éthiques généraux et mesures concrètes. Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie, 2025, vol. 64, no 1, p. 81-99.
3. Gouvernement du Québec. (1993). Loi sur la protection des renseignements personnels dans le secteur privé (mise à jour en 2020) L.Q., c. P-39.1, art. 12.1. https://www.legisquebec.gouv.qc.ca/fr/document/lc/p-39.1/.
4. Université de Montréal. (2018a). Déclaration de Montréal pour un développement responsable de l'intelligence artificielle. https://5da05b0d-f158-4af2-8b9f-892984c33739.filesusr.com/ugd/ebc3a3_28b2dfe7ee13479caaf820477de1b8bc.pdf/.

2. Overall Description

2.1 Product Perspective

This section is intentionally left blank for project-specific content. To be completed by the project team.

2.2 User Classes and Characteristics

2.2.1 Stakeholder Class and Priority

Table 2.1: Common stakeholder class in an AI project, their descriptions and priority based on the *Roles and Knowledge in Explainable AI (XAI)* framework.

#SC	Stakeholder Classes	Description	Priority
SC1	Primary	The people or groups who use the AI system and are directly impacted by its positive or negative results. In software development, the needs and expectations of primary stakeholders are often the most critical factors in design.	High
SC2	Secondary	People who work with the results produced by the system. Although the needs and expectations of secondary stakeholders are less critical than primary stakeholders, they can be just as influential and make it challenging to implement a new AI system within a company if their needs are not met.	Medium
SC3	Tertiary	People and groups affected more indirectly than secondary stakeholders. It is essential to involve tertiary stakeholders, as their opinions and perceptions can determine whether a project succeeds or fails	Low

** Priority level is determined based on stakeholder criticality and legal relevance.*

2.2.2 Stakeholders Descriptions

Table 2.2: Common stakeholders in AI projects and their descriptions, based on the *Roles and Knowledge in Explainable AI (XAI)* framework.

#S	Stakeholder	Description
Primary stakeholders		
S1	Party affected by the decision	Parties affected by the decision include all those impacted by the AI system. Certain aspects of their lives depend on the decision of an AI system.

Secondary stakeholders		
S2	User	Users include all those who take the recommendations issued by the AI system into consideration when making a decision.
Third party stakeholders		
S3	Data scientist	Data scientists develop AI algorithms and guide the understanding of AI systems from a scientific perspective.
S4	Developer	Developers implement AI systems and integrate them into the IT infrastructure.
S5	Project lead	Project leads are the people who decide where to employ specific systems. They manage the overall application domain in which AI systems operate. They do not necessarily interact directly with the AI system. Their decisions influence several other classes of stakeholders. They often act as intermediaries between several stakeholders.
S6	Manager	Managers manage a business, service, or administration.
S7	Ethicist	Ethicists explore the AI system's ethical, social and philosophical implications.
S8	Regulator	Regulators ensure compliance with the legal regulation of AI systems.
S9	Public	Not involved but aware of the project's existence.

2.2.3 Stakeholders Characteristics

Table 2.3: Knowledge types and their descriptions, based on the *Roles and Knowledge in Explainable AI (XAI)* framework.

K1. Knowledge in AI systems The knowledge required to research, develop, operate, or deploy AI systems.	
k1.a Technical knowledge in AI	Knowledge required to research and develop machine learning models.
k1.b Knowledge of AI system integrations	The knowledge required to operate and deploy AI systems.
K2. Knowledge of data The knowledge required to collect, organize, analyze, and communicate the data with which the model has been trained and makes decisions.	
K2.a Knowledge of data integration	The knowledge required for data collection and its preservation.
K2.b Knowledge of the domain expertise of data	The knowledge of the theory related to the domain expertise of the data and related technologies.

K3. Knowledge of the environment The knowledge of the environments in which human-AI interaction can occur.	
K3.a Knowledge of the physical environment	The knowledge of specific geographical locations or places where the AI system will be implemented and used.
K3.b Knowledge of the sociocultural environment	The knowledge of the set of events, facts, and phenomena, related to social, cultural, and economic aspects, observed in a studied physical environment.

Table 2.4: Degrees of knowledge and their descriptions, based on the *Roles and Knowledge in Explainable AI (XAI)* framework.

#	Degree	Formal knowledge	Instrumental knowledge
1	Advance	Yes	Yes
2	Basic	Yes	No
3	None	No	No

** Formal knowledge: An understanding of codified theories, embodied in texts or diagrams such as those found in textbooks, and is acquired over an extended educational process (Eraut, 2010).*
** Instrumental knowledge: An understanding of how to 'apply' formal knowledge. It is embodied in the use of tools or other instruments and is learned through demonstration and practice (Eraut, 2010).*

Table 2.5: Common degrees of knowledge for each stakeholder, by knowledge type, in AI projects, based on the *Roles and Knowledge in Explainable AI (XAI)* framework.

		K1 - AI		**K2 - Data**		**K3 - Env.**	
#S		**K1.a**	**K1.b**	**K2.a**	**K2.b**	**K3.a**	**K3.b**
SC1 - Primary stakeholders							
S1	Parties affected by the decision	3	3	3	3	3	3
SC2 - Secondary stakeholders							
S2	User	3	3	3	1	2	2
SC3 - Third party stakeholders							
S3	Data scientist	1	2	2	3	3	3
S4	Developer	2	1	1	3	3	3
S5	Project lead	2	2	2	2	1	1
S6	Manager	3	3	3	3	2	2

S7	Investor	3	3	3	3	3	3
S8	Ethicist	3	3	3	2	2	1
S9	Legislator	2	2	3	2	1	2
S10	Regulator	2	2	2	2	2	2
S11	Public	3	3	3	3	3	3

2.3 Operating Environment

No Template

2.4 Legal Constraints

Table 2.6: Key legal constraints for fully automated AI projects using personal data in Quebec. This table provides a non-exhaustive overview of the legal requirements to consider.

#C	Région	Loi	Art.	Description
C1	Quebec	Law 25	4	Any person carrying on a business who, by reason of a serious and legitimate interest, collects personal information about others must, prior to the collection, determine the purposes of the collection.
C2	Quebec	Law 25	3.3	Any person carrying on a business must carry out a privacy impact assessment for any project involving the acquisition, development or redesign of an information system or the electronic delivery of services involving the collection, use, disclosure, retention or destruction of personal information.
C3	Quebec	Law 25	8	The person who collects personal information from the person concerned must, at the time of collection and thereafter upon request, inform the person: (1) the purposes for which the information is collected; (2) the means by which the information is collected; (3) the rights of access and rectification provided by law; and (4) the right to withdraw consent to the communication or use of the information collected.
C4	Quebec	Law 25	3.2	Any person carrying on a business must establish and implement policies and practices governing its governance of personal information and ensuring its protection. Among other things, these policies and practices must provide a framework for the retention and destruction of personal information, define the roles and responsibilities of employees throughout

				the information life cycle, and establish a process for handling complaints about the protection of personal information. Detailed information about these policies and practices, particularly with regard to the content required by the first paragraph, is published in clear and simple terms on the company's website or, if the company does not have a website, is made available by any other appropriate means.
C5	Quebec	Law 25	12.1	Any person carrying on a business who uses personal information to make a decision based exclusively on the automated processing of such information must inform the person concerned no later than the time he or she informs him or her of the decision. They must also, at the request of the person concerned, inform him or her of: (1) the personal information used to make the decision; (2) the reasons, as well as the main factors and parameters, that led to the decision; (3) his or her right to have the personal information used to make the decision rectified. The person concerned must be given the opportunity to present his or her observations to a member of the company's staff who is in a position to review the decision.
C6	Quebec	Law 25	18.3	A person carrying on an enterprise may, without the consent of the person concerned, communicate personal information to any person or body if such communication is necessary for the exercise of a mandate or the performance of a contract for services or an undertaking entrusted to that person or body by the person carrying on the enterprise. In such a case, the person carrying on the business must: (1) give the mandate or contract in writing; (2) indicate, in the mandate or contract, the measures to be taken by the mandatary or person carrying on the contract to ensure that the confidentiality of the personal information communicated is protected, that the information is used only in the exercise of the mandate or the performance of the contract and that the information is not retained by the mandatary or person carrying on the contract after its expiry.
C7	Quebec	Law 25	23	Once the purposes for which personal information was collected or used have been fulfilled, the person carrying on the business must destroy or anonymize it in order to use it for serious and legitimate

				purposes, subject to a retention period stipulated by law.

** This table does not take into consideration the intended use of the AI nor the domain (ex. medical, banc, real estate). It is the organization's responsibility to add legal constraints regarding those aspects. Furthermore, beyond Quebec, at the federal level, Bill C-27 deals directly with AI systems. This legislation project follows in the footsteps of what is happening in Europe and the United States.*

2.5 Ethical Principles

Table 2.7: Common ethical principles applicable to XAI projects derived from the Montreal Declaration for a Responsible Development of Artificial Intelligence.

#P	Principle	Description regarding the implementation
P1	Respect for autonomy	It is crucial to empower citizens regarding digital technologies by ensuring access to the relevant forms of knowledge, promoting the learning of fundamental skills (digital and media literacy), and fostering the development of critical thinking.
P2	Solidarity	AI systems must be developed with the goal of collaborating with humans on complex tasks and should foster collaborative work between humans.
P3	Democratic participation	AI systems processes that make decisions affecting a person's life, quality of life, or reputation must be intelligible to their creators.
P4	Democratic participation	The decisions made by AI systems affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use. Justification consists in making transparent the most important factors and parameters shaping the decision, and should take the same form as the justification we would demand of a human making the same kind of decision.
P5	Democratic participation	Any person using a service should know if a decision concerning them or affecting them was made by an AI system.
P6	Responsibility	In all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decision must be taken by a human being and that decision should be free and informed.

2.6 User Documentation

This section is intentionally left blank for project-specific content. To be completed by the project team.

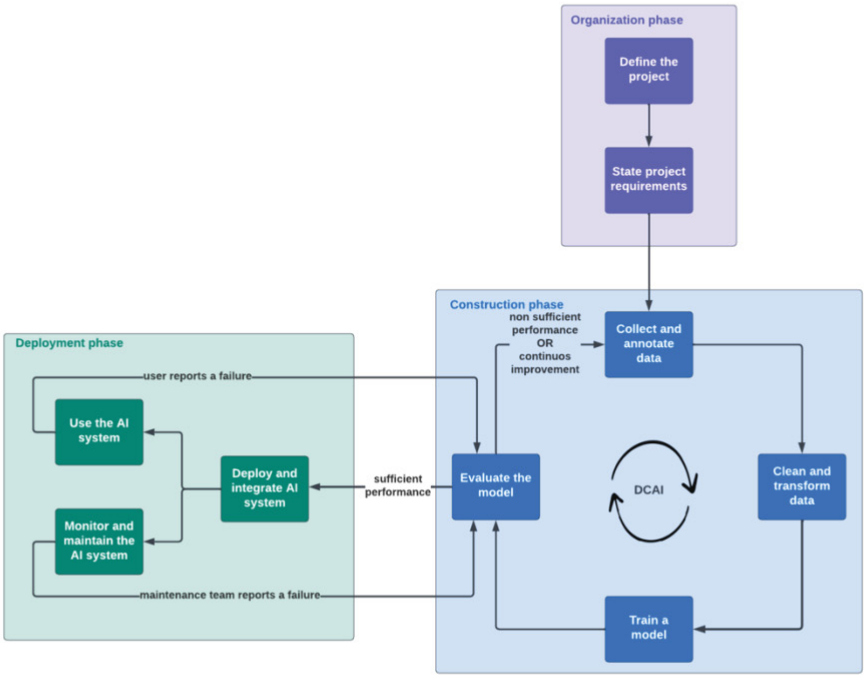
2.7 Assumptions and Dependencies

This section is intentionally left blank for project-specific content. To be completed by the project team.

3. System Interactions

3.1 System Life Cycle

Graph 3.1 : Common AI system life cycle using the Data-Centric AI process.



3.2 System Stakeholders Interactions

Table 3.1: Common stakeholders interactions with the AI system.

#I	Stakeholder	Interaction(s)
Use of the AI System		
I1	User	• Opt to utilize the AI system to obtain a prediction or recommendation. • Input the necessary data. • Request a prediction. • Analyze the results provided by the AI system. • Have confidence in the accuracy of the prediction. • Make a decision based on the results from the AI system and personal knowledge regarding the optimal use of this prediction. • Deliver the results to the affected party.

I2	Affected party	• Receive the prediction along with the user's decision. • Understand the prediction, including its value and underlying logic. • Learn how to modify or maintain the value of the prediction.
Monitor and maintain the AI system		
I3	User	• Identify inconsistencies in the results provided by the AI system. • Recognize that the logic employed by the algorithm is flawed. • Report the issue to the project lead.
I4	Data scientist	• Observe a decline in the performance of the predictions. • Report the issue to the project lead.
I5	Developer	• Ensure the system remains operational and that accurate data is consistently supplied to the AI through the pipeline. • Report any issues to the project lead.
I6	Project lead	• Analyze the reported issues. • Determine whether the system should return to the "Evaluate the model" phase for a deeper understanding of the errors or problems.
Deploy and integrate the AI system		
I7	User	• Receive training or a tutorial on how to operate the new AI system. • Test the system to better comprehend its capabilities.
I8	Data scientist	• Collaborate with the developer to ensure the AI system's algorithms are properly integrated into the environment. • Offer expertise on how the AI system should operate within the setting.
I9	Developer	• Implement the AI system within the existing IT infrastructure. • Resolve any technical issues that emerge during the deployment process. • Collaborate with the data scientist to verify that the system is functioning as intended.
I10	Project lead	• Oversee the deployment and integration process. • Ensure all stakeholders receive the necessary support and resources. • Monitor the initial performance and user acceptance of the deployed AI system. • Serve as the communication bridge between users, developers, and the manager.
I11	Regulator	• May require documentation or evidence of compliance with all relevant regulations and laws for the AI system's deployment.
I12	Ethicist	• Review the deployment of the AI system to confirm adherence to ethical guidelines.

I13	Public	• Needs to be informed about the deployment of AI systems as part of public transparency initiatives. • May be interested in the system's performance.
Evaluate the model		
I14	User	• Participate in validation studies to assess the effectiveness of the AI system. • Contribute expertise to refine the AI model based on the results of evaluations.
I15	Data scientist	• Analyze the AI's performance using specific metrics and benchmarks. • Adjust the algorithms and retrain the model based on performance feedback. • Document the evaluation process and results.
I16	Project lead	• Coordinate the evaluation process among all involved stakeholders. • Synthesize feedback and performance data to inform decisions on next steps. • Communicate the evaluation results and recommendations to the manager.
I17	Manager	• Review the evaluation report prepared by the project lead. • Make decisions regarding the future of the AI system within the organization based on this evaluation. • Ensure that the AI system continues to align with and meet organizational goals and requirements.
I18	Ethicist	• Review the evaluation results to ensure the AI system's use adheres to ethical standards. • Address any ethical concerns that arise from the evaluation data.
I19	Regulator	• May require evidence from the evaluation to ensure the AI system complies with regulatory standards.

3.3 Information Needs Assessment

Table 3.2: Common stakeholder information requirements during AI system lifecycle, and if XAI is not a available measure to gain this information.

#Inf	**Stakeholder**	**What (information needed)**	**When (lifecycle phase)**	**XAI Feat.**
Inf1	Affected Party	Prediction value definition	Use of the AI system	4.1
Inf2	Affected Party	How to change or conserve the prediction	Use of the AI system	4.4
Inf3	User	Data used by the AI system	Use of the AI system	4.1

Inf4	User	Prediction value definition	Use of the AI system	4.1
Inf5	User	How to change or conserve the prediction	Use of the AI system	4.4
Inf6	User	Confirmation that the prediction is good	Use of the AI system	4.5
Inf7	User	Confirmation that the explanation is good	Use of the AI system	4.6
Inf8	User	Prediction logical (local)	Use of the AI system	4.3
Inf9	User	AI system performance metrics	Monitor and maintain the AI system	4.1
Inf10	User	Prediction logic (local)	Monitor and maintain the AI system	4.3
Inf11	Data scientist	AI system performance metrics	Monitor and maintain the AI system	4.1
Inf12	Data scientist	XAI problem reports	Monitor and maintain the AI system	
Inf13	Developer	Operational status of the XAI components	Monitor and maintain the AI system	
Inf14	Developer	XAI problem reports	Monitor and maintain the AI system	
Inf15	Project lead	XAI problem reports	Monitor and maintain the AI system	
Inf16	User	AI logic (global)	Deploy and integrate the AI system	4.2
Inf17	User	Trained data	Deploy and integrate the AI system	4.1
Inf18	User	System's limits	Deploy and integrate the AI system	4.1
Inf19	User	Prediction value definition	Deploy and integrate the AI system	4.1
Inf20	User	AI system performance metrics	Deploy and integrate the AI system	4.1
Inf21	Project lead	AI logic (global)	Deploy and integrate the AI system	4.2
Inf22	Project lead	Trained data	Deploy and integrate the AI system	4.1
Inf23	Project lead	System's limits	Deploy and integrate the AI	4.1

			system	
Inf24	Project lead	Prediction value definition	Deploy and integrate the AI system	4.1
Inf25	Project lead	AI system performance metrics	Deploy and integrate the AI system	4.1
Inf26	Manager	AI logic (global)	Deploy and integrate the AI system	4.2
Inf27	Manager	AI system performance metrics	Deploy and integrate the AI system	4.1
Inf28	Public	The AI system has been deployed	Deploy and integrate the AI system	4.1
Inf29	Public	Ethical compliance evidence*	Deploy and integrate the AI system	
Inf30	Public	Regulatory compliance evidence*	Deploy and integrate the AI system	
Inf31	Public	AI system performance metrics	Deploy and integrate the AI system	4.1
Inf32	User	AI system performance metrics	Evaluate the model	4.1
Inf33	User	Prediction logic (local)	Evaluate the model	4.3
Inf34	User	AI logic (global)	Evaluate the model	4.2
Inf35	Data Scientist	AI system performance metrics	Evaluate the model	4.1
Inf36	Data Scientist	AI logic (global)	Evaluate the model	4.2
Inf37	Project lead	AI system performance metrics	Evaluate the model	4.1
Inf38	Project lead	AI error and limits analysis	Evaluate the model	
Inf39	Project lead	AI logic (global)	Evaluate the model	4.2
Inf40	Manager	AI system performance metrics	Evaluate the model	4.1
Inf41	Manager	AI logic (global)	Evaluate the model	4.2
Inf42	Ethicist	Ethical evaluation results*	Evaluate the model	
Inf43	Regulator	Regulatory compliance evidence*	Evaluate the model	
Inf44	Regulator	Regulatory compliance	Clean and transform data	

		evidence for data handling*		
Inf45	Regulator	Regulatory compliance evidence for data collection and annotation*	Collect and annotate data	

** Note: The information outlined in this table represents a high-level need. For a more comprehensive and detailed project plan, it would be ideal to further elaborate and subdivide these needs into more specific What (information needed). This will ensure a clearer definition of project needs, enabling more precise planning and execution. You can use the section 2.5 and 2.6 to help you create a more detailed What (information needed).*

4. Explainability Features

4.1 Provide a transparency note

4.1.1 Description and Priority

This feature provides clear information about the company's AI practices, including how the algorithm is developed, trained, and deployed, as well as their potential impact on users and society.
Stakeholders: Affected Party, User, Project lead, Manager, Public
Lifecycle phase: Use of the AI system, Monitor and maintain the AI system, Deploy and integrate the AI system, Evaluate the model
Priority: High

4.1.2 Functional Requirements

REQ1 - Definitions : This feature must provide the definitions of key attributes, including the predictions. It must present :

- Prediction value definition

REQ2 - Implementation choices: This feature must provide the stakeholders with choices made to develop the system. It should include :

- Training data description
- Impactful choices such as, but not limited to : how the data was handled, dataset division strategies, model optimisation strategies, etc
- Provide a high-level algorithmic architecture overview

REQ3 - Performance : This feature must provide the stakeholders with performance metrics regarding the system. It should include :

- Quantifiable metrics on the algorithms prediction reliability
- Known limits of the system

REQ4 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ4.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ4.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ4.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding

REQ5 - Simplified domain explanations for non-expert : The system must provide explanations that are easily understandable by a party who lacks knowledge in domain expertise.

REQ5.1 - Explanations must avoid domain expertise jargon and instead use clear, simple language that is accessible to non-domain stakeholders.

REQ5.2 - Incorporate a glossary or help section for terms and concepts related to the expertise domain, accessible directly within the explanation interface.

REQ5.3 - Where appropriate, use domain-expertise analogies or comparisons to familiar concepts to aid understanding.

REQ6 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ6.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ6.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ6.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ6.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ7 - Presentation format for non-domain expert stakeholders : The system must present domain expert concepts explanations in formats suitable for non-domain expert stakeholders.

REQ7.1 - Provide domain expert explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ7.2 - Use clear and simple graphics, where possible, to represent domain-specific concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ7.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the domain expert explanations.

REQ7.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific domain terminology.

REQ8 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

4.2 Provide global explanations

4.2.1 Description and Priority

Offer overarching insights into the decision-making processes of the AI model. It explains the general principles and mechanisms behind the model's decisions, enabling a comprehensive understanding of its functioning.

Stakeholders: User, Project lead, Manager, Data scientist

Lifecycle phase: Deploy and integrate the AI system, Evaluate the model

Priority: Medium

4.2.2 Functional Requirements

REQ1 - AI features : This functionality must provide the global explanations of the predictions. It must present :

- Prediction global explanations

REQ2 - Details : This feature must provide a level of detail sufficient for the understanding of the global logic of the model. It must provide, at least, those details:

- The 10 most impactful features in the dataset
- The general impact of those 10 features on the prediction
- How the influence of the top 5 features on the prediction changes based on their values

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.

REQ4 - Detailed domain explanations for domain-expert stakeholders: The explanations provided by the system must be tailored to reflect the expertise of the user, particularly focusing on the application in the field.

REQ4.1 - The explanations should integrate the domain terminology and concepts familiar to the professionals, especially those pertinent to their field.

REQ4.2 - The explanations can be clear, quick to read and to understand since the user already knows the important concepts regarding the prediction significance.

REQ5 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ5.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ5.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ5.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ6 - Presentation format for domain-expert stakeholders: The system must present domain expert explanations in formats suitable for domain-expert stakeholders.

REQ6.1 - Adapt the presentation of explanations to mimic traditional formats used in the specific domain of the stakeholders. It facilitates the integration of the system into their usual processes.

REQ7 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

4.3 Provide Local Explanations

4.3.1 Description and Priority

Provide detailed explanations for specific decisions or predictions made by the AI model. It allows users to understand the reasoning behind individual outputs, highlighting the factors and data points that influenced a particular decision.

Stakeholders: User

Lifecycle phase: Use of the AI system, Monitor and maintain the AI system, Evaluate the model

Priority: Medium

4.3.2 Functional Requirements

REQ1 - AI features : This functionality must provide the local explanations of some predictions. It must present :

- Prediction value definition and signification

REQ2 - Details : This feature must provide a level of detail sufficient for the understanding of the global logic of the model. It must provide, at least, those details:

- The 10 most impactful features in the dataset
- The impact of those 10 features on the prediction

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.

REQ4 - Detailed domain explanations for domain-expert stakeholders: The explanations provided by the system must be tailored to reflect the expertise of the user, particularly focusing on the application in the field.

REQ4.1 - The explanations should integrate the domain terminology and concepts familiar to the professionals, especially those pertinent to their field.

REQ4.2 - The explanations can be clear, quick to read and to understand since the user already knows the important concepts regarding the prediction significance.

REQ5 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ5.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ5.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ5.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ6 - Presentation format for domain-expert stakeholders: The system must present domain expert explanations in formats suitable for domain-expert stakeholders.

REQ6.1 - Adapt the presentation of explanations to mimic traditional formats used in the specific domain of the stakeholders. It facilitates the integration of the system into their usual processes.

REQ7 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

4.4 Provide ‘What Ifs’

4.4.1 Description and Priority

The ‘What If’ allows us to explore hypothetical scenarios by adjusting input variables and observing the impact on model predictions. This tool helps in understanding the model's behavior and decision-making process under different conditions.

Stakeholders: Affected Party, User

Lifecycle phases: Use of the AI system

Priority: High

4.4.2 Functional Requirements

REQ1 - AI features : This functionality must provide the ‘What Ifs’ of the prediction.

REQ2 - What ifs local logic: This feature must provide the new prediction’s value local logic.

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.

REQ4 - Simplified domain explanations for non-expert : The system must provide explanations that are easily understandable by a party who lacks knowledge in domain expertise.

REQ4.1 - Explanations must avoid domain expertise jargon and instead use clear, simple language that is accessible to non-domain stakeholders.

REQ4.2 - Incorporate a glossary or help section for terms and concepts related to the expertise domain, accessible directly within the explanation interface.

REQ4.3 - Where appropriate, use domain-expertise analogies or comparisons to familiar concepts to aid understanding.

REQ5 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ5.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ5.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ5.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ6 - Presentation format for non-domain expert stakeholders : The system must present domain expert concepts explanations in formats suitable for non-domain expert stakeholders.

REQ6.1 - Provide domain expert explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ6.2 - Use clear and simple graphics, where possible, to represent domain-specific concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ6.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the domain expert explanations.

REQ6.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific domain terminology.

REQ7 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

4.5 Provide metrics to assess the predictions trust

4.5.1 Description and Priority

Aims to enhance trust in AI predictions by providing robust metrics that assess the accuracy and reliability of a prediction. This functionality will deliver essential insights into the

confidence level of AI-generated outcomes, enabling stakeholders to make more informed decisions based on these predictions.
Stakeholders: User
Lifecycle phase: Use of the AI system
Priority: Medium

4.5.2 Functional Requirements

REQ1 - AI features : This functionality must provide confidence metrics for the prediction.
REQ2 - Confidence metrics: This feature must provide at least 2 confidence metrics:

- Provide input outlier detection
- Provide output outlier detection
- Add new metrics

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.

REQ4 - Detailed domain explanations for domain-expert stakeholders: The explanations provided by the system must be tailored to reflect the expertise of the user, particularly focusing on the application in the field.

REQ4.1 - The explanations should integrate the domain terminology and concepts familiar to the professionals, especially those pertinent to their field.

REQ4.2 - The explanations can be clear, quick to read and to understand since the user already knows the important concepts regarding the prediction significance.

REQ5 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ5.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ5.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ5.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ6 - Presentation format for domain-expert stakeholders: The system must present domain expert explanations in formats suitable for domain-expert stakeholders.

REQ6.1 - Adapt the presentation of explanations to mimic traditional formats used in the specific domain of the stakeholders. It facilitates the integration of the system into their usual processes.

REQ7 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

4.6 Provide metrics to assess the explanations' trust

4.6.1 Description and Priority

Focuses on quantifying the trustworthiness of these explanations. This feature introduces robust metrics to evaluate how faithful and consistent AI explanations are relative to the model's behavior.
Stakeholders: User
Lifecycle phases: Use of the AI system
Priority: Medium

4.6.2 Functional Requirements

REQ1 - AI features : This functionality must provide confidence metrics for the explanations.

- Global explanations
- Local explanations
- 'What If' explanations

REQ2 - Confidence metrics: This feature must provide at least 2 confidence metrics for this feature to be effective:

- Implement a faithfulness metric
- Implement a consistency metric

REQ3 - Simplified AI explanations for non-technical stakeholders: The system must provide AI explanations that are easily understandable by stakeholders who lack technical knowledge in AI and technology.

REQ3.1 - Explanations must avoid technical AI jargon and instead use clear, simple language that is accessible to non-technical stakeholders.

REQ3.2 - Incorporate a glossary or help section for terms and concepts related to AI, accessible directly within the explanation interface.

REQ3.3 - Where appropriate, use AI analogies or comparisons to familiar concepts to aid understanding.

REQ4 - Detailed domain explanations for domain-expert stakeholders: The explanations provided by the system must be tailored to reflect the expertise of the user, particularly focusing on the application in the field.

REQ4.1 - The explanations should integrate the domain terminology and concepts familiar to the professionals, especially those pertinent to their field.

REQ4.2 - The explanations can be clear, quick to read and to understand since the user already knows the important concepts regarding the prediction significance.

REQ5 - Presentation format for AI non-technical stakeholders : The system must present AI concepts explanations in formats suitable for non-technical stakeholders.

REQ5.1 - Provide AI explanations in multiple formats (text and visual) to cater to different user preferences and accessibility needs.

REQ5.2 - Use clear and simple graphics to represent AI concepts. These graphics should be accompanied by captions and explanations in simple language.

REQ5.3 - Use colors and icons to help visualize and distinguish different concepts or parts of the AI explanations.

REQ5.4 - Offer hierarchical or iterative views that allow follow-ups on initial explanations that use specific AI terminology.

REQ6 - Presentation format for domain-expert stakeholders: The system must present domain expert explanations in formats suitable for domain-expert stakeholders.

REQ6.1 - Adapt the presentation of explanations to mimic traditional formats used in the specific domain of the stakeholders. It facilitates the integration of the system into their usual processes.

REQ7 - Sensible data : The information distributed in this feature must not distribute internal or industrial secrets of the company.

Appendix A: Glossary

ANNEXE III

ÉVALUATION DE L'IMPACT DES CONTRIBUTIONS EN IA EXPLICABLE SUR UN PROJET INDUSTRIEL

Cette annexe introduit le document intitulé « Évaluation de l'impact des contributions en IA explicable sur un projet industriel ». Il s'agit d'un questionnaire soumis à des experts IA afin de recueillir leur analyse critique sur les retombées des livrables de cette thèse. Le document est présenté tel qu'il a été transmis aux participants.

Évaluation de l’impact des contributions de la présente thèse sur un projet industriel

L'objectif de ce sondage est de recueillir des données sur la pertinence et l'impact des contributions apportées par cette thèse dans le champ de l'intelligence artificielle explicable. En sollicitant le retour du responsable technique d'un projet de recherche industriel dans le domaine de la pharmaceutique, nous cherchons à comprendre comment les outils, méthodes et cadres proposés ont été perçus et utilisés dans un contexte pratique. Cette évaluation nous permettra de mesurer l'efficacité de nos propositions et d'identifier les éventuelles améliorations.

Ce sondage est structuré en sept sections distinctes. La section initiale vise à recueillir des informations sur le type de projet concerné et à dresser le profil du répondant. Les sections suivantes se penchent sur différents thèmes, correspondant chacun aux contributions spécifiques envisagées par cette thèse. La section finale élargit le champ d'exploration à une perspective globale, examinant non seulement l'impact du projet lui-même mais également son influence potentielle à l'échelle mondiale, en mettant l'accent sur l'efficacité des cadres, protocoles et méthodologies développés au cours de cette recherche.

Ce sondage inclut deux formats de questions. Le premier type comprend des questions à choix multiples, où vous êtes invité à évaluer votre réponse sur une échelle de 1 à 5, conformément aux descriptions fournies dans le tableau 1. Le second type est constitué de questions ouvertes à développement, vous offrant la liberté d'exprimer vos opinions et réflexions détaillées sur le sujet abordé.

Tableau 1 : Échelle d’évaluation de 1 à 5 pour les réponses du sondage.

1	2	3	4	5
Pas du tout	Peu	Suffisant	Beaucoup	Complètement

1. Définition du projet et du profil du répondant

1. Quel était votre rôle ou fonction dans la réalisation de ce projet ? Veuillez décrire vos responsabilités principales et fournir des exemples spécifiques de tâches pour illustrer votre contribution.

2. Dans quel secteur d'activité s'inscrivait votre projet d'IA ? Par exemple : finance, santé, droit.

3. L'explicabilité de l'IA (XIA) était-elle un élément critique pour la réussite de votre projet ? Veuillez expliquer en quoi cela a été le cas et discuter des impacts potentiels de la XIA sur l'aboutissement du projet.

4. Quelles limitations rencontrez-vous en matière d'explicabilité de l'IA (XIA) spécifiquement dans votre secteur d'activité ?

2. Protocole de définition des parties prenantes

5. Dans quelle mesure le protocole a-t-il été efficace pour identifier et catégoriser toutes les parties prenantes pertinentes de votre système de XIA ?

1	2	3	4	5	Non applicable

6. Le protocole a-t-il révélé des parties prenantes ou des intérêts que vous n'aviez pas initialement considérés ?

1	2	3	4	5	Non applicable

7. Dans quelle mesure ce protocole a-t-il facilité la gestion de votre projet en XIA ?

1	2	3	4	5	Non applicable

8. À partir de votre expérience avec le projet, pouvez-vous décrire l'impact global que le protocole de définition des parties prenantes a eu sur la réalisation de votre projet d'IA ? Merci de souligner comment ce protocole a influencé la dynamique de travail avec les parties prenantes, y compris tout impact sur la gestion du projet, la prise de décision, et l'atteinte des objectifs. Incluez également des réflexions sur les obstacles rencontrés et les leçons apprises qui pourraient guider les améliorations futures du protocole.

3. Définition des connaissances des parties prenantes

9. La méthodologie a-t-elle permis une évaluation précise des connaissances des parties prenantes en IA ?

1	2	3	4	5	Non applicable

10. Dans quelle mesure la méthodologie a-t-elle aidé à identifier les besoins en vulgarisation pour différentes parties prenantes ?

1	2	3	4	5	Non applicable

11. Avez-vous trouvé que les résultats de cette méthodologie facilitent la communication et la compréhension entre les développeurs de la XIA et les parties prenantes ?

1	2	3	4	5	Non applicable

12. Sur la base de votre interaction avec la méthodologie proposée pour évaluer les connaissances des parties prenantes dans le domaine de l'IA, pourriez-vous détailler comment cette approche a impacté votre capacité à comprendre et à répondre aux besoins de vulgarisation des différentes parties prenantes ? Veuillez partager des exemples concrets qui illustrent les succès rencontrés grâce à cette méthodologie, ainsi que les défis ou les limitations que vous avez observés.

4. Protocole de définition des besoins, exigences et contraintes en XIA

13. Dans quelle mesure le protocole a-t-il permis d'identifier de manière précise et complète les besoins, exigences, et contraintes spécifiques à votre système de XIA ?

1	2	3	4	5	Non applicable

14. Considérant le temps que vous avez investi pour appliquer le protocole de définition des besoins, exigences, et contraintes à votre système de XIA, estimez-vous que les résultats obtenus justifient cet investissement ?

1	2	3	4	5	Non applicable

15. Le protocole s'est-il avéré adaptable et flexible face à l'évolution des besoins et des contraintes au cours du projet ?

1	2	3	4	5	Non applicable

16. Comment l'utilisation de ce protocole a-t-elle affecté vos processus de prise de décision concernant le développement et l'implémentation de votre système de XIA ? A-t-elle conduit à des changements significatifs dans l'approche ou la stratégie du projet ? Veuillez partager des exemples concrets qui illustrent les succès rencontrés grâce à ce protocole, ainsi que les défis ou les limitations que vous avez observés.

5. Méthodologie de priorisation

17. La méthodologie de priorisation des besoins a-t-elle facilité la prise de décision lors du développement ou de l'implémentation de solutions XIA ?

1	2	3	4	5	Non applicable

18. De quelle manière cette méthodologie a-t-elle influencé la distribution des ressources, telles que le temps, le budget et le personnel, au sein de votre projet ?

1	2	3	4	5	Non applicable

19. La méthodologie de priorisation a-t-elle permis à votre projet de s'adapter rapidement aux changements ou aux nouveaux besoins qui ont émergé ?

1	2	3	4	5	Non applicable

20. En se concentrant sur les aspects prioritaires du projet identifiés par la méthodologie, avez-vous observé une amélioration de la qualité du système de XIA ou de l'innovation dans les solutions développées ? Comment ces améliorations peuvent-elles être attribuées à la méthodologie de priorisation utilisée ?

6. Intégration des impératifs légaux et éthiques

21. Dans quelle mesure l'intégration des impératifs légaux et éthiques dans le protocole vous a-t-elle encouragé à approfondir vos recherches sur les enjeux éthiques et légaux liés à l'IA ?

1	2	3	4	5	Non applicable

22. Dans quelle mesure l'intégration des impératifs légaux et éthiques a-t-elle amélioré la conformité de votre système de XIA ?

1	2	3	4	5	Non applicable

23. Le cadre proposé a-t-il facilité l'application des principes éthiques et légaux dans le développement de XIA ?

1	2	3	4	5	Non applicable

24. De quelle manière l'intégration des considérations légales et éthiques a-t-elle affecté la perception des parties prenantes externes (utilisateurs, clients, partenaires, financeurs) vis-à-vis de votre projet ? Avez-vous observé des retours ou des réactions spécifiques qui soulignent cette influence ?

7. Perception de l’apport global

25. Avez-vous déjà utilisé d'autres méthodologies pour définir les besoins, exigences, et contraintes de systèmes d'IA dans le passé ? Si oui, comment évalueriez-vous l'efficacité de ce protocole en comparaison ? Quels sont les avantages et inconvénients que vous avez observés ?

26. De quelle manière estimez-vous que les contributions de ces outils peuvent affecter l'expérience des utilisateurs finaux de systèmes d'IA ? Merci d'expliquer comment ces interactions pourraient se manifester dans l'utilisation quotidienne ou dans la perception de la technologie par le grand public.

27. De votre perspective, comment les outils et méthodologies proposés dans cette thèse peuvent-ils préparer les développeurs et les organisations à mieux faire face aux défis futurs de l'IA?

28. De votre perspective, de quelle manière les contributions de cette thèse peuvent-elles aider à renforcer la confiance des utilisateurs dans les technologies d'IA ?

29. De votre perspective, quel rôle les méthodologies proposées dans cette thèse pourraient-elles jouer dans l'influence des politiques publiques et des cadres législatifs relatifs à l'IA ?

BIBLIOGRAPHIE

Aahill, Nitinme, Urban, E., & Faley, P. (2023). Transparency Note for Azure AI Language. *Microsoft legal resources*. https://learn.microsoft.com/en-us/legal/cognitive-services/language-service/transparency-note/.

Accenture Conseils. (2018). *Intelligence artificielle, des conséquences réelles – Les services publics à l'ère de l'intelligence artificielle*. Accenture Conseils. https://www.accenture.com/acnmedia/PDF-88/Accenture-Artificial-Intelligence-Genuine-Impact.pdf.

Addario, F. e. S. S. (2022). *Validité constitutionnelle du projet de loi C-11, Loi sur la mise en œuvre de la Charte du numérique*. Commissariat à la protection de la vie privée du Canada.

Aholainen, M. (2025, May 13). *EU AI Act: Latest regulatory developments in Finland*. Hannes Snellman Attorneys Ltd. https://www.hannessnellman.com/news-and-views/blog/eu-ai-act-latest-regulatory-developments-in-finland/

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019, May). Fairwashing: the risk of rationalization. In *International Conference on Machine Learning* (pp. 161-170). PMLR.

Alvavi, Sepideh, E. S. C. & Perron, S. D. (2022). *Loi sur la protection de la vie privée des consommateurs du Canada (Projet de loi C-27) : incidences sur les entreprises*. Borden Ladner Gervais.

Amariles, D. R., & Baquero, P. M. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Review*, *48*, 105795.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-13).

Anders, C. J., Neumann, D., Samek, W., Müller, K. R., & Lapuschkin, S. (2023). Software for dataset-wide XAI: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy, 2021. *arXiv preprint arXiv:2106.13200*.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

Assemblée nationale du Québec. (2019). Étude détaillée du projet de loi n° 64, Loi modernisant des dispositions législatives en matière de protection des renseignements personnels. *Journal des débats de la Commission des institutions*, 42e législature, 1re session. https://www.assnat.qc.ca/fr/travaux-parlementaires/commissions/ci-42-1/journal-debats/CI-210527.html

Batic, D., Stankovic, V., & Stankovic, L. (2023). Towards transparent load disaggregation–a framework for quantitative evaluation of explainability using explainable AI. *IEEE Transactions on Consumer Electronics*.

Belaid, M. K., Hüllermeier, E., Rabus, M., & Krestel, R. (2022). Do we need another explainable AI method? Toward unifying post-hoc XAI evaluation methods into an interactive and multi-dimensional benchmark. *arXiv preprint arXiv:2207.14160*.

Benchekroun, O., Rahimi, A., Zhang, Q., & Kodliuk, T. (2020). The need for standardized explainability. *arXiv preprint arXiv:2010.11273*.

Bérubé, N. (2023). Peut-on vraiment contrôler l'IA ? *La Presse*. https://www.lapresse.ca.

Braun, M., Bleher, H., & Hummel, P. (2021). A leap of faith: is there a formula for "Trustworthy" AI? *Hastings Center Report*, *51*(3), 17-22.

Brie, E. (2018). Beyond culture: Geographic relocation and social trust across Canada. *Unpublished manuscript*.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Castro, C., O'Brien, D., & Schwan, B. (2023). Egalitarian machine learning. *Res Publica*, *29*(2), 237-264.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*.

Charoliya, W. (2023). Data-Centric AI: The Key to Unlocking the Full Potential of Machine Learning. *Codemotion Magazine*. https://www.codemotion.com/magazine/ai-ml/data-centric-ai-the-key-to-unlocking-the-full-potential-of-machine-learning/.

Chiaroni, J., Zillner, S., Bertels, N., Bezombes, P., Bonhomme, Y., Amadou-Boubacar, H., ... & Zissis, D. (2021). Franco-German position paper on" Speeding up industrial AI and trustworthiness".

Chromik, M., & Butz, A. (2021). Human-XAI interaction: a review and design principles for explanation user interfaces. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18* (pp. 619-640). Springer International Publishing.

Crockett, K., Colyer, E., Gerber, L., & Latham, A. (2021). Building trustworthy AI solutions: A case for practical solutions for small businesses. *IEEE Transactions on Artificial Intelligence*, *4*(4), 778-791.

Commission de l'éthique en science et en technologie (2021). *Les effets de l'intelligence artificielle sur le monde du travail et la justice sociale*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2023a). *La gestion algorithmique de la main-d'œuvre : analyse des enjeux éthiques*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2023b). *La transformation numérique du réseau de la santé et des services sociaux en vue d'intégrer l'intelligence artificielle : un regard éthique*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2023c). *Mériter et renforcer la confiance des citoyens dans la gestion et la valorisation des données de santé : pour une gouvernance transparente et responsable, soucieuse de la dignité des personnes et de l'intérêt public*. Gouvernement du Québec.

Commission de l'éthique en science et en technologie (2024). *Intelligence artificielle générative en enseignement supérieur : enjeux pédagogiques et éthiques*. Gouvernement du Québec.

Commission européenne. (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Commission européenne, Direction générale des réseaux de communication, du contenu et des technologies, (2019). *Lignes directrices en matière d'éthique pour une IA digne de confiance*, Publications Office. https://data.europa.eu/doi/10.2759/74304

Conseil Européen. (2016a). Article 15, Droit d'accès de la personne concernée. Dans *Règlement général sur la protection des données*.

Conseil Européen. (2016b). Article 22, Décision individuelle automatisée, y compris le profilage. Dans *Règlement général sur la protection des données*.

Conseil Européen. (2016c). Considérant 71. *Règlement général sur la protection des données*.

Council, N. R. (1999). *Funding a Revolution: Government Support for Computing Research*. Washington, DC: The National Academies Press. https://doi.org/10.17226/6323.

Denis, P. L., Morin, D., & Guindon, C. (2010). Exploring the capacity of NEO PI-R facets to predict job performance in two French-Canadian samples. *International Journal of Selection and Assessment*, *18*(2), 201-207.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (pp. 1591-1602).

Direction de l'accès à l'information et de la protection des renseignements personnels, Secrétariat à la réforme des institutions démocratiques et à l'accès à l'information. (2010). *Guide de références : Processus d'évaluation des risques d'atteinte à la protection des renseignements personnels liés aux projets d'acquisition, de développement et de refonte d'un système d'information ou de prestation électronique de services*.

Eason, K. D. (1989). *Information technology and organisational change*. CRC Press.

Eraut, M. (2010). Knowledge, working practices, and learning. *Learning through practice: Models, traditions, orientations and approaches*, 37-58.

Felzmann, H., Fosch-Villaronga, E., Lutz, C. et A. Tamò-Larrieux (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361. https://doi.org/10.1007/ s11948-020-00276-4

Ferlitsch, Andrew. (2023). Making the machine: The machine learning lifecycle. *Google Cloud Blog*. https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle?hl=en

Ferguson, A., Munjal, A et M. Hartmann (2024). Use AI to Check Your Message-Content Spam Score. *Microsoft Customer Insights*. https://learn.microsoft.com/en-us/dynamics365/customer-insights/journe ys/spam-checker

Ferrer, X., Van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72-80.

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, 19(4), 534–560. https://doi.org/ 10.1017/epi.2022.39

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185-193.

Gagnon, F. & Casanova, C. (2023). Pour un débat rationnel sur l'intelligence artificielle. *La Presse*. https://www.lapresse.ca.

Garrett, N., Beard, N. et C. Fiesler (2020). More Than "If Time Allows": The Role of Ethics in AI Education. Dans A. Markham, J. Powles, T. Walsh et A. L. Washington (dir.), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (p. 272-278)*. Association for Computing Machinery.

Gaudin, J. & Élisabeth Lesage-Bigras. (2019). Droit à l'explication de l'intelligence artificielle - un regard transatlantique. *ROBIC*. https://www.robic.ca/publications/droit-a-lexplication-de-lintelligence-artificielle-un-regard-transatlantique/.

Goffi, E. (2023). Les robots tueurs sont inévitables. *Le Devoir*. https://www.ledevoir.com/opinion/idees/525221/les-robots-tueurs-sont-inevitables.

Gouvernement du Canada. (2018). Principe 3 - Fournir des explications claires. *Principes directeurs en matière d'IA*.

Gouvernement du Canada (2020). *Projet de loi C-11 - Loi édictant la Loi sur la protection de la vie privée des consommateurs et la Loi sur le Tribunal de la protection des renseignements personnels et des données et apportant des modifications corrélatives et connexes à d'autres lois.* https://www.parl.ca/legisinfo/fr/projet-de-loi/43-2/c-11.

Gouvernement du Canada (2022a). Article 11(2). *Projet de loi C-27 - Loi édictant la Loi sur la protection de la vie privée des consommateurs, la Loi sur le Tribunal de la protection des renseignements personnels et des données et la Loi sur l'intelligence artificielle et les données et apportant des modifications corrélatives et connexes à d'autres lois.* https://www.justice.gc.ca/fra/sjc-csj/pl/charte-charter/c27_1.html.

Gouvernement du Canada (2022b). Article 2. *Projet de loi C-27 - Loi édictant la Loi sur la protection de la vie privée des consommateurs, la Loi sur le Tribunal de la protection des renseignements personnels et des données et la Loi sur l'intelligence artificielle et les données et apportant des modifications corrélatives et connexes à d'autres lois*. https://www.justice.gc.ca/fra/sjc-csj/pl/charte-charter/c27_1.html.

Gouvernement du Canada (2022c). Article 63. *Projet de loi C-27 - Loi édictant la Loi sur la protection de la vie privée des consommateurs, la Loi sur le Tribunal de la protection des renseignements personnels et des données et la Loi sur l'intelligence artificielle et les données et apportant des modifications corrélatives et connexes à d'autres lois*. https://www.justice.gc.ca/fra/sjc-csj/pl/charte-charter/c27_1.html.

Gouvernement du Canada (2022d). *Projet de loi C-27 - Loi édictant la Loi sur la protection de la vie privée des consommateurs, la Loi sur le Tribunal de la protection des renseignements personnels et des données et la Loi sur l'intelligence artificielle et les données et apportant des modifications corrélatives et connexes à d'autres lois*. https://www.justice.gc.ca/fra/sjc-csj/pl/charte-charter/c27_1.html.

Gouvernement du Canada. (2023a). *Évaluation de l'incidence algorithmique (EIA)*. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/evaluation-incidence-algorithmique.html

Gouvernement du Canada. (2023b). *Guide sur l'utilisation de l'intelligence artificielle générative*. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/guide-utilisation-intelligence-artificielle-generative.html

Gouvernement du Canada. (2024). *Liste des fournisseurs d'intelligence artificielle (IA) intéressés*. https://www.canada.ca/fr/gouvernement/systeme/gouvernement-numerique/innovations-gouvernementales-numeriques/utilisation-responsable-ai/liste-fournisseurs-intelligence-artificielle-ia-interesses.html

Gouvernement du Québec. (1993). *Loi sur la protection des renseignements personnels dans le secteur privé* (mise à jour en 2020) L.Q., c. P-39.1, art. 12.1. https://www.legisquebec.gouv.qc.ca/fr/document/lc/p-39.1/.

Gouvernement du Québec (2021a). Article 12.1. *Loi 64 - Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*. https://www.assnat.qc.ca/fr/travaux-parlementaires/projets-loi/projet-loi-64-42-1.html?appelant=MC.

Gouvernement du Québec (2021b). Stratégie d'intégration de l'intelligence artificielle dans l'administration publique 2021 à 2026. *Bibliothèque et Archives nationales du Québec*. https://www.quebec.ca/gouvernement/politiques-orientations/vitrine-numeriqc/strategie-integration-ia-administration-publique-2021-2026.

Gouvernement du Québec. (2021c). *Loi modernisant des dispositions législatives en matière de protection des renseignements personnels*. L.Q., c. 25.

Gouvernement du Québec (2022). *Espace évolutif – Modernisation des lois: Tout savoir en continu sur la modernisation des lois sur la protection des renseignements personnels au Québec*. https://www.cai.gouv.qc.ca/espace-evolutif-modernisation-lois/.

Gowling WLG & IAB Canada. (2023). *Rapport du sondage sur la Loi 25 : les organisations sont-elles prêtes pour la nouvelle loi québécoise sur la protection de la vie privée*. Gowling WLG. Récupéré de https://gowlingwlg.com/getmedia/22f5bf42-ab9e-468f-82aa-2c9d86ad284c/Gowling-WLG-et-IAB-Canada_Rapport-du-sondage-sur-la-Loi-25.pdf.xml/.

Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face Recognition Vendor Test (FRVT) Part 3: Demographic effects* (NISTIR 8280). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8280

Hamid, O. H. (2022). From model-centric to data-centric AI: A paradigm shift or rather a complementary approach?. In *2022 8th International Conference on Information Technology Trends (ITT)* (pp. 196-199). IEEE.

Haque, A. B., Islam, A. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120.

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). sage.

Hofstede Insights. (2023). *Country Comparison Tool*. https://www.hofstede-insights.com/country-comparison-tool/.

Hussain, F., Hussain, R., & Hossain, E. (2021). Explainable artificial intelligence (XAI): An engineering perspective. *arXiv preprint arXiv:2101.03613*.

IEEE (1984). IEEE Guide for Software Requirements Specifications. IEEE Std 830-1984, 1-26. https://doi.org/10.1109/IEEESTD.1984.119205/.

IEEE. (2021). *7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. IEEE Standards Association. https://standards.ieee.org/standard/7000-2021.html

Innovation, Sciences et Développement économique Canada. (2021). Opinions des Canadiens sur l'intelligence artificielle. *Publications du gouvernement du Canada*. https://publications.gc.ca/collections/collection_2021/isde-ised/Iu4-396-2021-fra.pdf.

Innovation, Sciences et Développement économique Canada. (2019). Renforcer la protection de la vie privée dans l'ère numérique - Propositions pour moderniser la Loi sur la protection des renseignements personnels et des documents électroniques. *Publications du gouvernement du Canada*. https://ised-isde.canada.ca/site/innover-meilleur-canada/fr/charte-canadienne-numerique/renforcer-protection-vie-privee-dans-lere-numerique.

Intact Assurance (s. d.). Termes et conditions de l'application Mon Intact. https://www.intact.ca/fr/termesapplication-intact#maconduite-v4

International Organization for Standardization. (2021). *ISO/IEC TR 24029-1: Intelligence artificielle (IA) : Évaluation de la robustesse des réseaux de neurones*. https://www.iso.org/standard/77609.html/.

ISO/IEC JTC 1/SC 42. (2020). *Artificial intelligence — Overview of trustworthiness in artificial intelligence* (ISO/IEC TR 24028:2020). International Organization for Standardization.

ISO/IEC JTC 1/SC 42. (2022). *Artificial intelligence — Artificial intelligence concepts and terminology* (ISO/IEC TR 22989:2022). International Organization for Standardization.

ISO/IEC JTC 1/SC 42. (2023). *Artificial intelligence — Guidance on risk management* (ISO/IEC TR 23894:2023). International Organization for Standardization.

Ishizaki, Kazuaki. (2020). AI Model Lifecycle Management: Overview. *IBM Blog*. https://www.ibm.com/blog/ai-model-lifecycle-management-overview/.

Jang, Y., Kim, K., Leite, F., Ayer, S., & Cho, Y. K. (2021). Identifying the perception differences of emerging construction-related technologies between industry and academia to enable high levels of collaboration. *Journal of Construction Engineering and Management*, *147*(10), 06021004.

Joncas, H. (2023). Nouvelles règles sur les renseignements personnels : Des PME à la traîne. La Presse. https://www.lapresse.ca/affaires/entreprises/2023-09-11/nouvelles-regles-sur-les-renseignements-personnels/des-pme-a-la-traine.php/.

Jyoti, A., Ganesh, K. B., Gayala, M., Tunuguntla, N. L., Kamath, S., & Balasubramanian, V. N. (2022). On the robustness of explanations of deep neural network models: A survey. *arXiv preprint arXiv:2211.04780*.

Knight, W. (2016). Tesla Crash Will Shape the Future of Automated Cars. *Technology Review*. https://www.technologyreview.com/2016/07/01/70693/tesla-crash-will-shape-the-future-of-automated-cars/.

Kostick-Quenet, K., Lang, B. H., Smith, J., Hurley, M., & Blumenthal-Barby, J. (2023). Trust criteria for artificial intelligence in health: normative and epistemic considerations. *Journal of Medical Ethics*.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

Larsson, S. (2019). The socio-legal relevance of artificial intelligence. *Droit et société*, *103*(3), 573-593.

Larue Langlois, R. (2023). Le CRIM appuiera un consortium de recherche industriel sur l'IA de confiance. *Direction informatique*. https://www.directioninformatique.com/le-crim-appuiera-un-consortium-de-recherche-industriel-sur-lia-de-confiance/15/10/2023/.

Le, P. Q., Nauta, M., Van Bach Nguyen, S. P., Pathak, S., Schlötterer, J., & Seifert, C. (2023, August). Benchmarking eXplainable AI-A Survey on Available Toolkits and Open Challenges. *IJCAI*, 6665-6673.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of management Review*, *23*(3), 438-458.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social forces*, *63*(4), 967-985.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, *4*(8), 669-677.

Liao, Q. V., Subramonyam, H., Wang, J. et J. Wortman Vaughan (2023). Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-powered User Experience. Dans *Proceedings of the 2023 CHI conference on human factors in computing systems* (p. 1-21).

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Luhmann, N. (1979). Trust and power (H. Davis, J. Raffan, & K. Rooney, Trans.). *UK: John Wiley and Sons*.

MacMillan, D. (2025, January 13). *Arrested by AI: Police ignore standards after facial recognition matches*. The Washington Post. https://www.washingtonpost.com/business/interactive/2025/police-artificial-intelligence-facial-recognition/

Mahima, K. Y., Ayoob, M., & Poravi, G. (2021). An Assessment of Robustness for Adversarial Attacks and Physical Distortions on Image Classification using Explainable AI. In *AI-Cybersec@ SGAI* (pp. 14-28).

Majeed, A., & Hwang, S. O. (2023). Technical Analysis of Data-Centric and Model-Centric Artificial Intelligence. *IT Professional*, *25*(6), 62-70.

Marcellis-Warin, N., Dostie, B., Dufour, G., Armellini, F., Aubert, B. A., Beaudry, C., ... & Zhegu, M. (2021). *Le Québec économique 9: Perspectives et défis de la transformation numérique* (No. 2020li-01). CIRANO.

Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. *Vintage*.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

McKenna, A. (2021). La protection de la vie privée au Québec a désormais beaucoup de mordant. *Le Devoir*. https://www.ledevoir.com/economie/635412/la-protection-de-la-vie-privee-au-quebec-a-desormais-beaucoup-de-mordant/.

McKenna, A. (2023). Québec respecte mal sa propre loi sur les renseignements personnels. *Le Devoir*. https://www.ledevoir.com/societe/798778/cybersecurite-quebec-respecte-mal-propre-loi-renseignements-personnels/.

Memarian, B. et T. Doleck (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and Higher Education: A Systematic Review. *Computers and Education: Artificial Intelligence*, 5, 100–152. https://doi.org/10.1016/j.caeai.2023.100152

Middleton, S. E., Letouzé, E., Hossaini, A., & Chapman, A. (2022). Trust, regulation, and human-in-the-loop AI: within the European region. *Communications of the ACM*, *65*(4), 64-68.

Mila. (2020). *Rapport annuel : Du 1er avril 2019 au 31 mars 2020*. Mila. https://mila.quebec/wp-content/uploads/2020/12/Mila-Rapport-Annuel-2020-Web.pdf/.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-45.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869-877.

Nadeau, J.-B. (2022). Le Québec aux avant-postes en matière de protection des données personnelles. *Le Devoir*. https://www.ledevoir.com/societe/693318/protection-des-donnees-le-quebec-aux-avant-postes/.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1-42.

Neely, M., Schouten, S. F., Bleeker, M. J., & Lucic, A. (2021). Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*.

Netflix (s. d.). How to Rate TV Shows and Movies? *Netflix Help Center*. https://help.netflix.com/en/node/ 9898

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Ng, A. (2021). A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI. *DeepLearningAI*. https://www.youtube.com/watch?v=06-AZXmwHjot=1607s/.

Ngan, M., Grother, P., & Hanaoka, K. (2020). Ongoing Face Recognition Vendor Test (FRVT) Part 6A: *Face recognition accuracy with masks using pre-COVID-19 algorithms* (NISTIR 8311). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8311

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics and information technology*, 24(1), 13.

Office of the Superintendent of Financial Institutions Canada. (2023). *Financial Industry Forum on Artificial Intelligence: A Canadia Perspective on Responsible AI*. https://www.osfi-bsif.gc.ca/sites/default/files/documents/ai-ia_en.pdf

Pandianchery, M. S., & Ravi, V. (2022). Explainable AI framework for COVID-19 prediction in different provinces of India. *arXiv preprint arXiv:2201.06997*.

Parthasarathy, V., Urban, E., Faley, P. (2023). Transparency Note for Image Analysis. *Microsoft legal resources*. https://learn.microsoft.com/en-us/legal/cognitive-services/computer-vision/imageanalysis-transparency-note/.

Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15, Article 31652. https://doi.org/10.1038/s41598-025-17242-4

Piorkowski, D., Park, S., Wang, A. Y., Wang, D., Muller, M., & Portnoy, F. (2021). How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-25.

Price Waterhouse Coopers. (2022). Analyse économique des investissements réalisés en intelligence artificielle au Québec - Rapport détaillé. Forum IA Québec. https://api.forumia.devbeet.com/app/uploads/2022/03/pwc_forumiaqc_sommaire_2022.pdf/.

Public Works and Government Services Canada. (1997). *The Canadian style: A guide to writing and editing: Revised and expanded*. Dundurn Press in cooperation with Public Works and Government Services Canada, Translation Bureau. https://publications.gc.ca/pub?id=9.646585&sl=1

Raz, A. K., Nolan, S. M., Levin, W., Mall, K., Mia, A., Mockus, L., ... & Williams, K. (2022, March). Test and evaluation of reinforcement learning via robustness testing and explainable ai for high-speed aerospace vehicles. *2022 IEEE Aerospace Conference (AERO)* (pp. 1-14). IEEE.

Rebstadt, J., Remark, F., Fukas, P., Meier, P., & Thomas, O. (2022). Towards personalized explanations for AI systems: designing a role model for explainable AI in auditing.

Reddy, G. P., & Kumar, Y. P. (2023, April). Explainable ai (xai): Explained. *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)* (pp. 1-6). IEEE.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749-2767.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.

Sartor, G., & Lagioia, F. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.

Schuchmann, S. (2019). Analyzing the prospect of an approaching AI winter. *Unpublished doctoral dissertation*.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.

Statistics Canada. (2017). *Census in brief: English, French, and official language minorities in Canada*. Statistics Canada. https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016011/98-200-x2016011-eng.cfm/.

Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Tobey, A. (2019). *Explainability: Where AI and liability meet*. ARIAS. https://www.arias-us.org/wp-content/uploads/2019/04/Tobey-Explainability-Where-AI-and-Liability-Meet.pdf

Tobin, J., Le, J., & Rachakonda, V. (2022). Lecture 1: Course vision and when to use ML. *Full Stack Deep Learning*. https://fullstackdeeplearning.com/course/2022/lecture-1-course-vision-and-when-to-use-ml/.

United States Senate. (2022). S.3572 - Algorithmic Accountability Act of 2022. Congress.gov. Retrieved from https://www.congress.gov/bill/117th-congress/senate-bill/3572

Université de Montréal. (2018a). *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*. https://5da05b0d-f158-4af2-8b9f-892984c33739.filesusr.com/ugd/ebc3a3_28b2dfe7ee13479caaf820477de1b8bc.pdf/.

Université de Montréal. (2018b). Principe 5 : Principe de participation démocratique. *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*.

van Wynsberghe, A. (2016). *Healthcare Robots: Ethics, Design and Implementation*. Routledge.

van Berkel, N., Tag, B., Goncalves, J. et S. Hosio (2022). Human-Centred Artificial Intelligence: A Contextual Morality Perspective. *Behaviour & Information Technology*, 41(3), 502–518. https://doi.org/ 10.1080/0144929x.2020.1818828

Vanhala, M., Heilmann, P., & Salminen, H. (2016). Organizational trust dimensions as antecedents of organizational commitment. *Knowledge and Process Management*, *23*(1), 46-61.

Vitrine IA Québec. (n.d.). Le Québec regorge d'innovateurs et d'entrepreneurs en IA. https://vitrine.ia.quebec/repertoire/region-de-quebec/.

Weber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, 92, 154-176.

Weller, A. J. (2019). Design thinking for a user-centered approach to artificial intelligence. *She Ji: The Journal of Design, Economics, and Innovation*, *5*(4), 394-396.

White & Case. (2024). *AI Watch: Global regulatory tracker – Spain*. White & Case LLP. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-spain

XGBoost Developers. (2022). *XGBoost Parameters*. XGBoost. https://xgboost.readthedocs.io

Yu, R., & Shi, L. (2018). A user-based taxonomy for deep learning visualization. *Visual Informatics*, *2*(3), 147-154.

Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., & Hu, X. (2023). Data-centric ai: Perspectives and challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)* (pp. 945-948). Society for Industrial and Applied Mathematics.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., ... & Zumar, C. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull., 41*(4), 39-45.

Zhang, J. (2022). *User Interface Design Based on Human-Centered Explainable AI Methods* (Master's thesis).

Zhang, M. (2015). Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software. *Forbes*. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

Zhou, J., Gandomi, A. H., Chen, F. et A. Holzinger (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5), 593. https://doi.org/10.3390/electroni cs10050593