N°6 | STEPWISE

IA DE CONFIANCE

DU CONCEPT A L'ACTION !

Quantmetry

Executive Summary

EXECUTIVE SUMMARY

L’intelligence artificielle (IA) est aujourd’hui un moteur essentiel du développement économique. Le développement inéluctable de l’IA comme outil de productivité présente cependant, selon ses domaines d’application, des risques de différentes natures : sécurité et santé, socio-économique, éthique, juridique, etc. En outre, certains risques sont difficilement perceptibles ; ainsi par exemple l’usage de l’IA pour le scoring ou l’aide à la sélection, peut présenter de lourdes conséquences éthiques, s’il se fonde sur des données d’entraînement involontairement biaisées.

La nécessité de maîtriser de tels risques conduit la Commission Européenne à renforcer le cadre réglementaire qui entoure les usages les plus critiques du Big Data et de l’IA. Cet enjeu de normalisation s’inscrit dans une stratégie visant à donner toute sa place à l’Europe dans la compétition internationale.

Notre conviction est qu’un cadrage normatif, contraignant dans les seules situations à risque, aura également des implications positives sur les IA non critiques. En effet, il est probable qu’en élevant les attentes et exigences de qualité, l'Europe parvienne à imposer progressivement de nouveaux standards et de nouvelles pratiques, comme on a pu l’observer avec la RGPD.

Par ailleurs, la réglementation peut opportunément servir de cadre de référence aux entreprises pour qu’une IA construite suivant ces principes soit aussi source de valeur :

• Par une efficacité collective accrue : la confiance dans la façon dont les outils utilisant de l’IA ont été conçus et intégrés aux processus métier facilite leur acceptation et leur bonne exploitation par les utilisateurs au sein de l’entreprise.
• Par la confiance de ses clients ou usagers dans la fiabilité de ses produits et services et la compréhension des décisions prises,
• Par une plus grande robustesse : des résultats plus fiables et plus stables dans le temps.

Selon l’enquête que nous avons réalisée avec 25 entreprises ayant des solutions IA industrialisées, une réglementation d’IA de confiance telle que la Commission Européenne l’envisage, est perçue plutôt comme une nécessité que comme un frein. Cependant, sa définition et surtout sa mise en œuvre restent encore incertaines pour la plus grande partie des participants.

Dans ce livre blanc, nous vous proposons une démarche de mise en œuvre afin de pouvoir s’inscrire dans la vision et les exigences futures d’une IA de confiance en précisant des moyens techniques et organisationnels et en donnant des exemples concrets de cas d’usage et de méthodes appliqués.

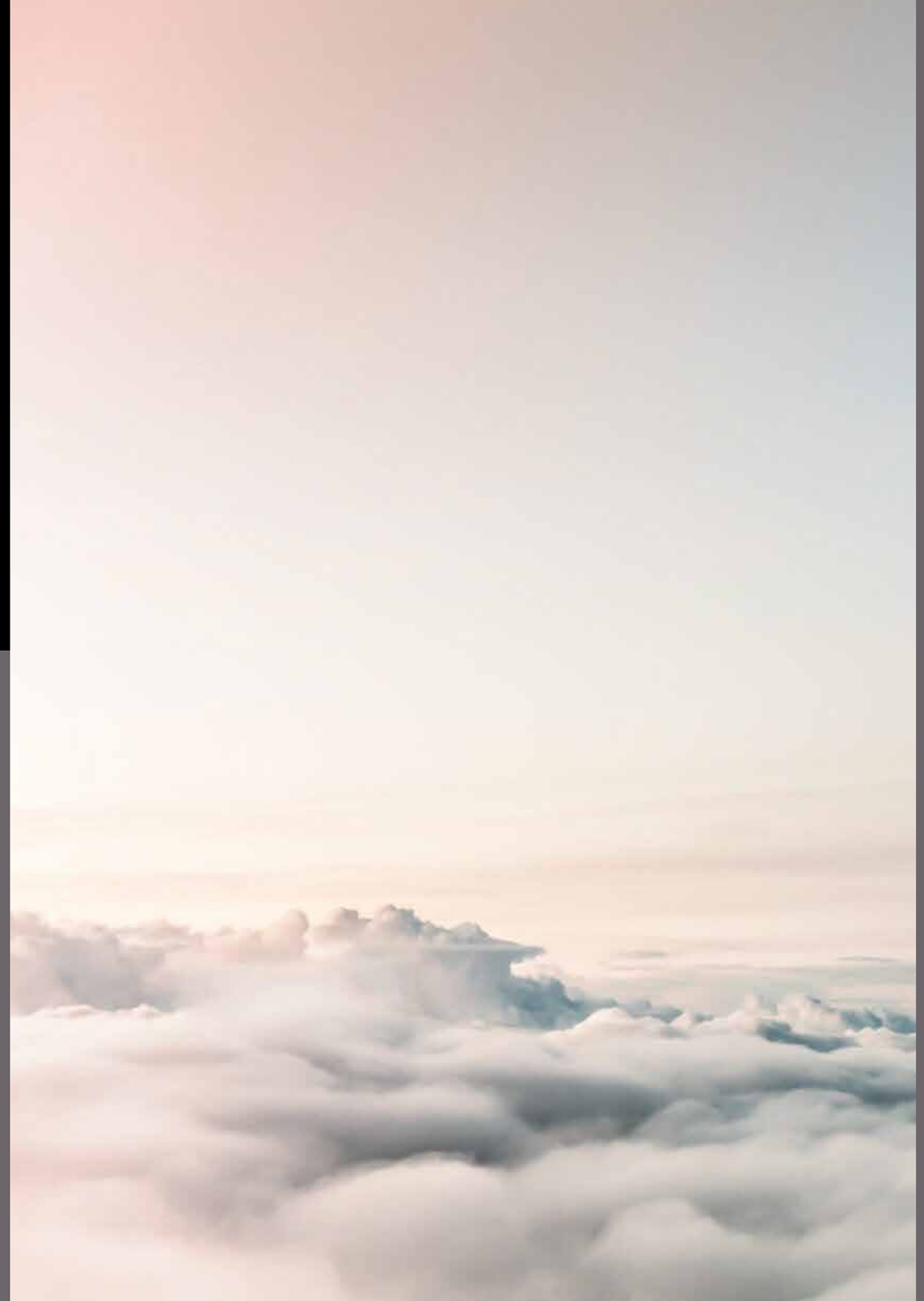
Les avancées récentes et les perspectives de la recherche dans les domaines de l'IA éthique, robuste et intelligible, concluent notre livre blanc.

Guillaume Bodiou
Partner chez Quantmetry

Quantmetry

The State of the Art AI company.

TABLE DES MATIÈRES



Préface

De nombreuses grandes entreprises, institutions de recherche, institutions publiques nationales et internationales ont publié leurs principes d'IA de confiance (également qualifiée d'IA responsable, ou encore éthique). Plusieurs principes, telle l'explicabilité au cœur des travaux de l'ACPR en gouvernance de l'IA et qui faisait l'objet d'un précédent livre blanc de Quantmetry, se retrouvent dans la plupart de ces chartes. Leur plus grand dénominateur commun correspond essentiellement aux 7 principes de l'IA éthique énoncés par la Commission Européenne et rappelés dans ce qui suit.

Or si ces chartes permettent de compléter les obligations réglementaires par des préoccupations sociales ou éthiques, les principes qu'elles énoncent sont souvent difficilement traduisibles en objectifs concrets et atteignables au sein des processus métier.

Une bonne pratique, indispensable pour relever ce défi, consiste à définir et prendre en compte les objectifs éthiques dès la conception d'un projet, notamment afin de gérer les attentes : l'IA ne peut pas tout, doit être encadrée, et ses objectifs sont souvent en tension réciproque, imposant par exemple comme les travaux de l'ACPR l'ont illustré, de sacrifier quelques points de performance pour satisfaire une contrainte de traitement équitable par l'algorithme.

Un obstacle supplémentaire provient du nombre de parties prenantes à chaque objectif de l'IA responsable. La confiance doit en effet être établie auprès de l'ensemble des utilisateurs directs ou indirects, et plus largement des individus ou groupes potentiellement affectés par les décisions prises ou facilitées par l'IA. Par ailleurs – comme le montre l'enquête réalisée par Quantmetry et décrite dans ces pages – la responsabilité du suivi des principes de l'IA de confiance s'avère difficile à assigner tant les fonctions impliquées sont nombreuses.

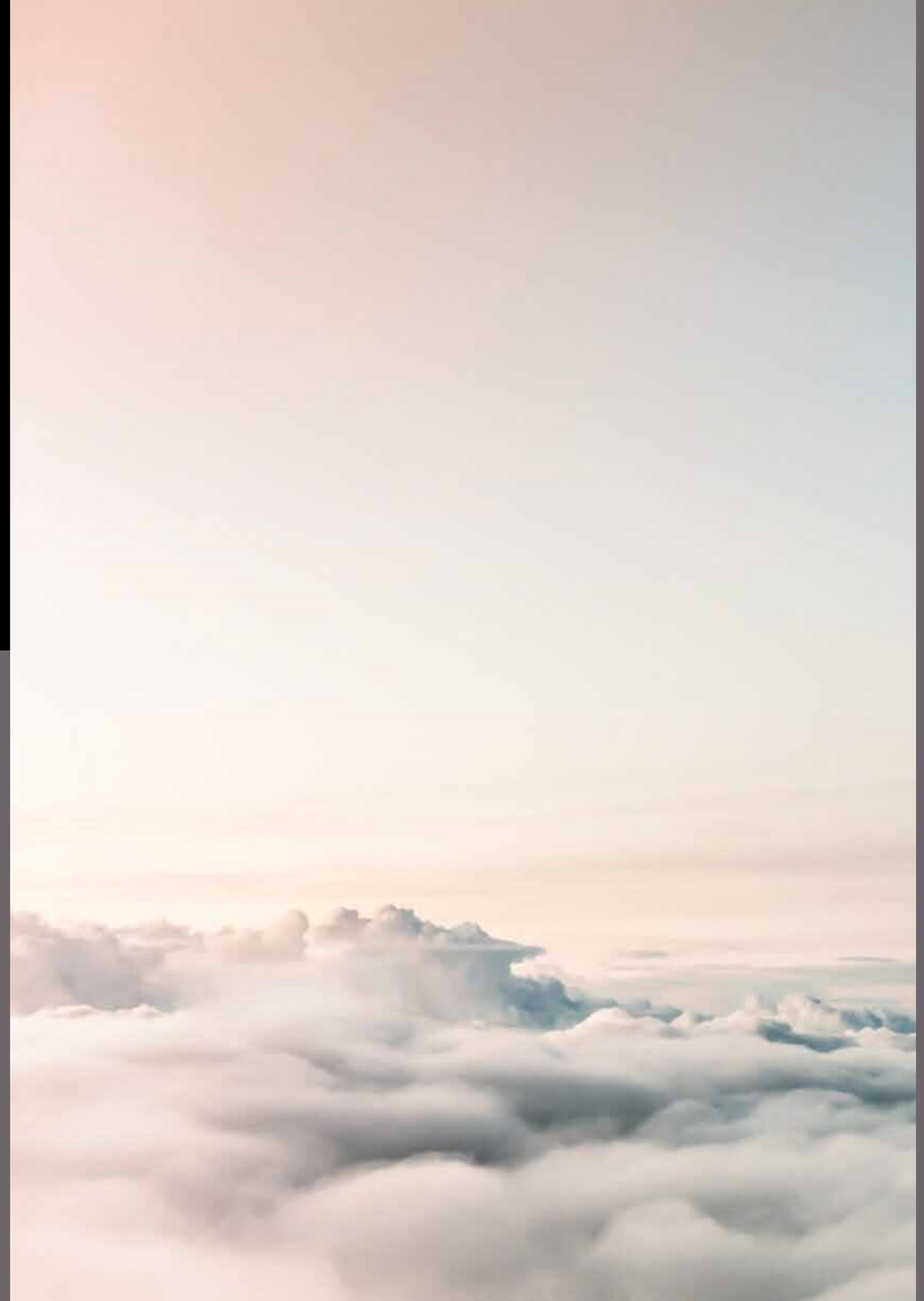
En outre, poser des principes est une étape nécessaire mais non suffisante à la production d'IA responsable. Il conviendra aussi de définir une méthodologie d'audit appropriée, permettant de détecter les risques associés, d'y remédier, de vérifier la conformité réglementaire, et plus généralement de mettre en musique les principes préalablement définis dans une organisation.

Parallèlement devraient voir le jour des normes et standards internationaux, visant à produire des critères homogènes de conformité des systèmes à base d'IA.

Ainsi, les méthodes, processus et outils propres à soutenir une IA de confiance nécessitent encore une phase de maturation dans la plupart des organisations. Cette démarche itérative est inhérente à la complexité de certains algorithmes mis en œuvre – complexité décuplée par celle de leur intégration dans les systèmes traditionnels (legacy systems) qui prévalent encore dans les secteurs assurantiel et bancaire. Elle provient aussi de la difficile orchestration des partitions jouées par les différents acteurs de l'adoption d'IA : compréhension commune d'enjeux de nature pluridisciplinaire, de la part d'équipes dont la collaboration est souvent inhibée par une séparation fonctionnelle entre experts techniques et métier. Une IA de confiance est avant tout une IA responsable, au double sens de responsibility et accountability. Sa réalisation exigera donc, au-delà de simplement « cocher la case » dans une liste de principes à respecter, la mise en place d'un ensemble de bonnes pratiques de conception et de contrôle de l'IA (incluant les piliers que sont l'explicabilité et l'équité algorithmique, afin comme l'explique ce livre blanc « d'établir puis entretenir la confiance ») et une culture organisationnelle ayant assimilé ces sujets.

En contrepartie de ces nouveaux attendus réglementaires et organisationnels, et comme ce fut le cas par exemple en matière de protection des données personnelles et de la vie privée suite à l'adoption du RGPD, de nouveaux modèles et opportunités d'affaires devraient émerger afin de gérer les risques associés à l'IA et d'obtenir le niveau de confiance escompté : offres de conseil, services et produits d'audit voire de certification, etc. Autant de façons de cueillir les fruits d'une innovation réfléchie et concertée.

Laurent Dupont, ACPR
Autorité de Contrôle
Prudentiel et de Résolution



Remerciements

REMERCIEMENTS

Merci aux 25 entreprises interrogées et à leurs dirigeants pour la confiance qu'ils nous ont accordée, ainsi que pour la qualité de nos échanges, notamment :
Arkea, CACIB, Covea, EDF Groupe, La Française des jeux (FDJ), Groupama, LNE, Malakoff Humanis, Michelin, SwissLife, TheFork, Suez.

Merci à l'ACPR pour avoir accepté de préfacer ce livre blanc et pour la qualité des travaux menés sur le thème de l'IA de confiance qui profitent à l'ensemble de la communauté.

Merci aux rédacteurs et aux relecteurs de ce livre blanc : Anne Champalone, Guillaume Bodiou, Martin Le Loc, Nicolas Brunel, Amélie Segard, Tsvetina Bacheva, Maya Azouri, Paul Brenot-Vinciguerra, Philippe Neveux, Youssef Horchani, Grégoire Martinon, Vlad Flamind, Guillaume Besson, Victor Gouin.

Introduction

L'IA est généralement définie comme un ensemble de concepts et de technologies mises en œuvre en vue de réaliser des systèmes capables de reproduire le comportement humain. Ce livre blanc s'intéresse plus particulièrement à l'IA non symbolique, en l'occurrence le Machine Learning, que l'on peut définir comme l'ensemble des algorithmes capables d'apprendre à résoudre un problème, depuis les données, sans être explicitement programmés.

La confiance peut être définie comme le sentiment de sécurité et de sûreté qu'a une personne vis-à-vis de quelqu'un ou de quelque chose. Pour faire confiance, il faut pouvoir croire à l'autre et accepter le fait d'être dans un état de vulnérabilité. En effet, accorder sa confiance c'est donner à l'autre partie un certain pouvoir sur nous. Elle est primordiale car sans confiance, on ne pourrait même pas envisager l'avenir et chercher à bâtir un projet qui se développe dans le temps. Pour Emmanuel Kant, pas de doute, la question de confiance est fondamentale. Il disait dans la Métaphysique des mœurs « Dans un monde où la confiance n'existe pas, les devoirs de loyauté tombent en désuétude ». À la lumière de cette définition, nous allons commencer par caractériser ce qu'est une IA de confiance.

L'ère de l'IA gagne chaque jour un peu plus en maturité et se développe rapidement en tant que technologie dont les applications semblent illimitées. Désormais, les technologies de l'IA deviennent de plus en plus omniprésentes, responsables d'un nombre croissant de décisions importantes qui impactent directement la vie humaine. Par conséquent, les utilisateurs se montrent davantage concernés par leurs relations avec les algorithmes. C'est pour cela qu'une relation de confiance entre l'être humain et l'IA est devenue indispensable. En effet, une IA de confiance permet d'une part aux utilisateurs de se sentir en sécurité dans son utilisation et ainsi de les rassurer, favorisant l'adhésion des citoyens à l'égard des algorithmes d'IA. D'autre part, elle permet à l'entreprise de maîtriser toujours plus son usage pour en maximiser les gains d'application mais aussi pour éviter des scandales éthiques de plus en plus fréquents ayant un impact direct sur l'image de l'entreprise concernée. Pour parvenir à cela, des efforts délibérés doivent être faits afin de mettre en place les bonnes structures de gouvernance au sein des entreprises et de sensibiliser l'ensemble de l'organisation aux principes de l'IA de confiance.

Fort de ce constat, et étant donné l'importance que revêt le sujet de l'IA, la Commission Européenne a décidé d'investir à partir de 2020, 20 milliards d'euros dans l'IA chaque année contre 3,3 milliards d'euros pour l'année 2016 (Vakulina, 2019). Elle a aussi formé depuis Avril 2018 un groupe d'experts qui travaille notamment à l'élaboration du cadre réglementaire de l'IA dont la sortie est imminente. En effet, ce cadre réglementaire global au niveau de l'Union Européenne (UE) permettrait non seulement de protéger les utilisateurs de l'IA mais aussi de construire un avantage différenciant pour les acteurs européens.

Ce travail de l'UE pour protéger ses citoyens s'inscrit dans la continuité des travaux engagés dès 2018 dans le domaine du respect de la protection des données personnelles et des droits fondamentaux des citoyens face aux défis du Big Data. Effectivement, avec la mise en place du RGPD, un premier pas dans la direction d'une IA de confiance a été fait en déployant un premier niveau de défense pour les citoyens Européens. Ce dernier doit maintenant être consolidé avec l'apparition du cadre réglementaire de l'IA.

01

Un impératif sociétal : s’engager à identifier et maîtriser les risques du déploiement de l’IA

Des risques émergents, amplifiés, voire insoupçonnés

L'IA est aujourd'hui au cœur d'une transformation significative de notre société. Nombreuses de ses applications ont un impact économique notable, en améliorant par exemple les rendements industriels par l'optimisation des chaînes de production ou encore en maximisant l'effet d'une campagne promotionnelle grâce au marketing hyper-personnalisé. D'autres cas d'usage de l'IA tels que l'aide au diagnostic en santé ou l'amélioration des flux de mobilité en ville participent quant à eux à améliorer notre bien-être sociétal et environnemental. L'IA est de plus en plus présente dans nos vies professionnelles comme personnelles.

Ce développement inéluctable de l'IA présente cependant, selon ses domaines d'application, des risques de différentes natures. Dans ce livre blanc, nous mettrons l'emphase sur trois de leurs aspects : leur caractère émergent, leur amplification par les mécanismes de biais et enfin leur caractère parfois insidieux.

Des risques émergents

De plus en plus de cas d'usage de l'IA sont fondés sur des modèles algorithmiques dont la complexité et le niveau d'abstraction dépassent les limites de l'intelligibilité par l'humain, entraînant de façon significative la diminution de la compréhension et donc de sa capacité à maîtriser le système en question.

Ces cas d'usages ne se limitent pas à des tests en laboratoires mais sont progressivement mis en œuvre dans notre quotidien, pour notre confort par exemple dans le cas des aides à la conduite déployées par les constructeurs automobiles dans nos voitures, mais parfois indépendamment de notre volonté, comme les algorithmes de recommandation qui influent sur notre consommation.

Nous assistons ainsi progressivement à la diminution de l'intervention humaine dans des domaines impactant notre vie quotidienne où toute défaillance dans la décision du modèle algorithmique pourrait avoir des conséquences hasardeuses. Suivant l'expansion de l'usage de l'IA, ces risques vont continuer à se multiplier : ils convient donc de les comprendre pour mieux les maîtriser.

Des risques amplifiés

Comme nous le savons, le cœur des algorithmes de machine learning est le jeu de données d'entraînement. Si ce dernier est constitué de données présentant un biais statistique, il existe un risque non négligeable que le modèle d'IA entraîné reproduise, voire amplifie ce biais.

En effet, un modèle apprend son comportement des données historiques, même si cela n'est pas toujours souhaitable. Les données peuvent refléter un déséquilibre entre sous-populations parce que le passé est biaisé sous l'effet de pratiques discriminatoires historiques qui infusent dans les données, ou encore sous l'effet d'une collecte inadéquate des données d'apprentissage (mauvaise représentativité des populations par exemple).

Le modèle d'IA aboutit alors à des décisions biaisées qui se renforcent par ces biais de sélection s'ils ne sont pas détectés, mesurés et traités tout au long du traitement de la donnée.

Les algorithmes de recommandations utilisées aujourd'hui par de nombreuses entreprises illustrent parfaitement ces risques amplifiés. En effet, sur les plateformes de streaming musical par exemple, les algorithmes peuvent mettre en lumière des morceaux de musique qui, sans ces modèles algorithmiques, passeraient inaperçus.

Analysant l'assortiment de sentiments suscité par le morceau, accompagné d'un traitement de données personnelles, le modèle algorithmique met en évidence des similarités avec d'autres morceaux écoutés par d'autres utilisateurs et le considère ainsi comme un morceau recommandable. Par conséquent, un véritable effet boule de neige se produit. Plus le morceau est recommandé aux auditeurs, plus il est écouté et ainsi plus il est recommandé (Carpentier, 2021). Finalement, une chanson qui est passée inaperçue à l'origine devient connue et écoutée par tous.

Des risques parfois insoupçonnés

S'il y a des domaines d'application et des cas d'usage en IA dont on perçoit facilement la dimension critique, tels que les véhicules autonomes ou les dispositifs médicaux intelligents, il y en a d'autres où cette criticité n'est pas immédiatement perceptible, notamment dans le domaine des interactions sociales ou des jugements faisant appel à des valeurs morales. L'IA étant finalement, à l'image de ses concepteurs, soumise aux différents biais auxquels notre société nous expose, et cela souvent de manière imperceptible, elle peut en tirer les pires travers.

Le tristement célèbre algorithme de Google dont le but initial était de classifier des images en est la parfaite illustration. En 2015, l'algorithme a classifié des images d'hommes Afro-américains avec le label "Gorilla", déclenchant une polémique évidente et justifiée. La raison principale de cette confusion provient d'un déséquilibre dansles jeux de données d'entraînement des algorithmes de reconnaissance faciale qui sont majoritairement composés de personnes de couleur blanche, rendant l'algorithme moins pertinent à l'égard des personnes noires. Ce biais de représentation dans la population d'entraînement qui n'a pas du tout été soupçonné s'est avéré ainsi être problématique.

La maîtrise de tels risques insoupçonnés à la conception est une des raisons pour lesquelles la Commission Européenne envisage de renforcer le cadre réglementaire qui entoure les usages liés au Big Data et à l'IA. De fait, nous ne pouvons pas nous passer d'IA pour certains usages, mêmes critiques. Il en est ainsi par exemple de l'affectation via Parcoursup des étudiants en France, qui est d'une complexité telle qu'elle doit nécessairement faire appel à des algorithmes non interprétables par les citoyens.

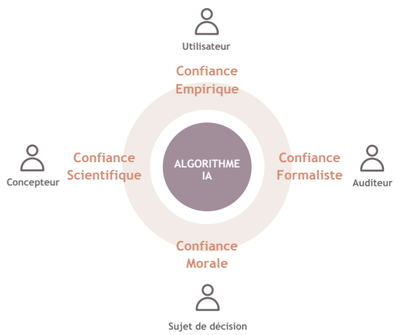
Les risques liés à l'usage de l'IA sont donc déjà présents et nous l'avons vu, protéiformes. Face à cette diversité, une prise de conscience est nécessaire de la part de l'ensemble des acteurs et des utilisateurs passifs de l'IA pour pouvoir mettre en œuvre les techniques permettant de gagner la confiance nécessaire à une IA transformatrice de notre société.

Des besoins de confiance diversifiés

Nous avons présenté dans l'introduction de ce livre blanc quelques éléments de définition de la confiance et mis en lumière la difficulté de la définir de manière unique. Cette variété de définitions s'explique particulièrement dans le cadre de l'IA par l'hétérogénéité des acteurs qui contribuent à la conception, l'implémentation, l'utilisation et le contrôle des algorithmes.

L'utilisation de plus en plus fréquente d'algorithmes d'IA fait naître plusieurs besoins, parfois contradictoires. En s'intéressant aux interactions entre les rôles intervenant tout au long du cycle de vie de l'algorithme, nous pourrons mieux identifier et décrire leurs attentes en termes de confiance.

Figure 1 : Rôles principaux autour d'un algorithme d'IA



Nous distinguerons quatre rôles principaux composant les interactions majeures autour d'un algorithme d'IA :

- **L'utilisateur**, qui interagit directement avec l'algorithme, en renseignant les données et en collectant les résultats ;
- **Le sujet de décision**, qui est directement affecté par la décision assistée par l'IA ;
- **Le concepteur**, qui a développé et mis en production l'algorithme ;
- **L'auditeur**, qui est en charge d'évaluer le modèle et de vérifier la conformité.

Chacun de ces rôles s'accompagne d'objectifs et de préoccupations propres, desquelles découle une certaine vision de ce que représente la confiance en l'algorithme. Qui plus est, tous ces acteurs ne sont pas forcément des personnes physiques, mais peuvent représenter des personnes morales, des entreprises ou même des niveaux plus abstraits d'IA.

L'utilisateur souhaitera obtenir un résultat issu de l'algorithme qui sera à la fois le plus efficace pour répondre à la problématique posée, mais également qui ne soit pas trop éloigné de celui qu'il aurait, par ses connaissances ou par son expérience, lui-même donné. La confiance dans l'algorithme est pour lui principalement basée sur des critères empiriques.

Le sujet de décision (notamment le citoyen) est principalement intéressé par son propre cas et évaluera la confiance à l'échelle de l'individu selon des critères proches de la morale. Il souhaite être traité de façon équitable, dans son individualité, et comprendre les facteurs qui ont joué en sa faveur ou défaveur. Ce rôle fait émerger une contrainte forte d'individualité dans les attentes de confiance.

Le concepteur (le plus souvent un Data Scientist) souhaitera concevoir et implémenter l'algorithme le plus performant et robuste et recherchera donc une vision quasi-mathématique de la confiance. Celle-ci se traduisant par une valeur suffisamment satisfaisante d'une ou plusieurs métriques de performance définies lors de la conception.

L'auditeur a pour mission de contrôler le système afin de garantir un certain niveau de performance sur des critères parfois différents de ceux du concepteur, voire le cas échéant de procéder à des enquêtes pour identifier des dysfonctionnements. Il adopte une vision formaliste de la confiance, en cherchant à l'évaluer selon les critères définis par le standard ou la législation.

Il apparaît clairement que ces acteurs n'ont pas les mêmes attentes en matière de confiance, ni n'en ont une définition unique. A un niveau plus global, on peut s'interroger sur l'aspect culturel de la confiance : l'Europe semble à cet égard se détacher du reste du monde comme une figure de proue de la défense des intérêts des citoyens en voulant faire de l'IA de confiance un des principes fondamentaux de sa stratégie IA, quand les États-Unis d'une part affichent une importance moins marquée et donc une législation plus souple à ce sujet, et la Chine d'autre part déploie un système de crédit social opaque et sans droit à la contestation. L'IA de confiance apparaît donc protéiforme et se décline selon les craintes, les ambitions et la culture de chaque individu ou régime étatique.

Une Europe réglementée pour une IA de confiance

Comme abordé précédemment, l'UE fait figure de leader mondial sur le sujet de l'IA de confiance en s'étant emparé relativement tôt du sujet et en se dotant progressivement d'un cadre légal adapté aux enjeux de l'exploitation de la donnée (directive ePrivacy, GDPR, Data Governance Act), dynamique qui s'est accélérée depuis le début de l'année 2020, marquée notamment par la sortie du livre blanc « Intelligence Artificielle - Une approche européenne axée sur l'excellence et la confiance ».

Figure 2 : Calendrier des initiatives réglementaires Data & IA de l'Union européenne

L'UE a ainsi mis en place une approche basée sur la gestion des risques visant à développer une IA éthique, porteuse des valeurs européennes et protégeant les citoyens européens, et à éviter l'apparition de réglementations nationales hétérogènes complexes à appliquer pour les acteurs économiques paneuropéens.

Une approche fondée sur 7 exigences clés afin de promouvoir une IA éthique, respectant les droits de l'homme

Depuis juin 2018, la Commission Européenne a ainsi réuni un groupe d'experts, AI HLEG (High Level Expert Group), afin de travailler sur le cadrage de l'IA de confiance et proposer des recommandations sur le futur cadre réglementaire de l'UE attendu pour 2021. Avec ce groupe d'experts, la Commission Européenne a défini 7 exigences auxquelles les IA doivent répondre afin de préserver les droits fondamentaux.

Figure 3 : Les 7 exigences pour une IA de confiance

Le groupe d'experts a également proposé, en juin 2020, une grille d'évaluation pour une IA digne de confiance (ALTAI) fondée sur ces 7 exigences clés ainsi qu'un outil d'auto-évaluation à destination des acteurs de l'IA. Cet outil d'auto-évaluation – à compléter avec l'ensemble de parties prenantes intervenant dans le développement de vos IA – aide à mieux comprendre les différents critères d'une IA de confiance et à prendre conscience de la diversité des risques de dérive d'une IA. Cette méthode permet d'obtenir un score de confiance dans les algorithmes et des recommandations afin de s'orienter dès à présent vers la mise en œuvre d'une IA de confiance. Le groupe d'expert a de plus proposé des méthodes techniques et non techniques afin de mettre en œuvre les 7 exigences. La section 3 de ce livre blanc, proposant des orientations relatives à la mise en place d'une IA de confiance, est dans la lignée de ces recommandations.

Une approche par la criticité afin de maintenir l'équilibre entre confiance et innovation

Enfin et surtout, le 20 octobre 2020, le parlement européen a adopté à une grande majorité un texte sur la réglementation de l'IA. Ce texte confirme que la réglementation sur l'IA de confiance sera fondée sur les 7 exigences clés et que celle-ci s'appliquera uniquement aux IA dites « critiques ». En effet, le souci de soutenir les entreprises dans la mise en place de leur IA de confiance en évitant de brider l'innovation, conduit la Commission Européenne à légiférer non pas sur les technologies elles-mêmes mais sur les domaines d'application où l'IA est critique. Pour les IA non critiques, la Commission se limite à inciter à suivre ses recommandations.

Figure 4 : Liste des secteurs et des usages à risque selon la Commission Européenne

Secteurs à haut risque

Secteur public (asile, migration, contrôles aux frontières, système judiciaire et services de sécurité sociale)
Défense et sécurité
Finance, banque et assurance
Emploi
Éducation
Soins de santé
Transports
Énergie

Usages ou finalités à haut risque

Recrutement
Notation et évaluation des étudiants
Affectation de fonds publics
Octroi de prêts
Commerce, courtage, fiscalité, etc.
Traitements et procédures médicaux
Processus électoraux et campagnes politiques
Gestion des déchets
Décisions du secteur public ayant une incidence importante et directe sur les droits et obligations des personnes physiques ou morales
Conduite automatisée
Gestion du trafic
Systèmes militaires autonomes
Production et distribution d'énergie
Contrôle des émissions

Liste exhaustive et cumulative des secteurs à haut risque et des usages ou finalités à haut risque comportant un risque de porter atteinte aux droits fondamentaux et aux règles de sécurité.

Des entreprises s’engagent dans la mise en œuvre d’une IA de confiance reconnue (Enquête Quantmetry)

Objectif & Cadre de l’enquête

En février-mars 2021, nous avons réalisé une enquête auprès d’un panel d’entreprises utilisant des algorithmes d’IA pour répondre aux questions suivantes :

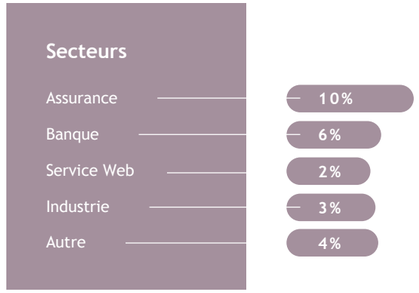
• Quel est leur avis par rapport au projet réglementaire de la Commission Européenne sur l’IA de confiance ?
• Quelles sont les craintes principales liées à cette nouvelle réglementation ?
• Au-delà de la réglementation, ces entreprises ont-elles déjà mis en place des mesures d’IA de confiance ? Si oui, lesquelles et pourquoi ?
• Quelles sont leurs trois prochaines priorités en lien avec l’IA de confiance ?

Synthèse des résultats de l’enquête

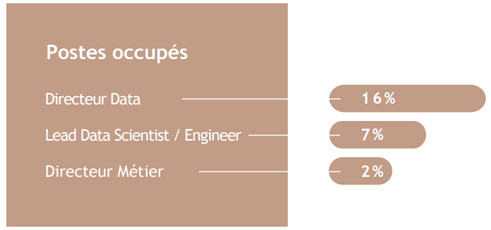
25 interviews d’acteurs IA :

Dans le cadre de la nouvelle réglementation relative à l’IA de confiance, nous avons interrogé un panel de 25 entreprises allant de l’ETI au CAC40, principalement dans le secteur bancaire et assurantiel, afin de mieux comprendre la manière dont ils anticipent et préparent sa mise en œuvre.

Secteurs couverts

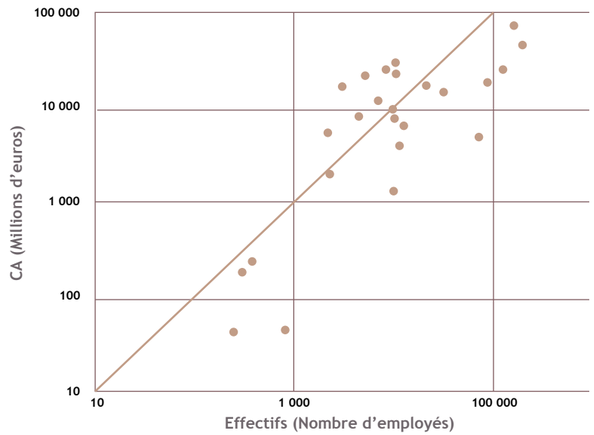


Postes occupés / métiers



Distributions CA vs. Effectifs

Distribution des entreprises par CA/EFFECTIFS



Synthèse des résultats de l’enquête

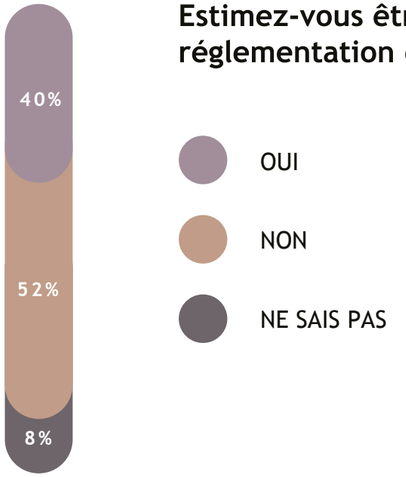
Chiffres clés de l’enquête :

- **100%** des entreprises consultées ont au moins un système IA industrialisé
- **80%** ont mis en place des contrôles d’intelligibilité des algorithmes IA
- **80%** perçoivent une réglementation d’IA de confiance comme une nécessité
- **40%** estiment être suffisamment informés sur la future réglementation européenne de l’IA de confiance
- **63%** considèrent que certains cas d’usage seront concernés par la future réglementation européenne de l’IA de confiance
- **64%** anticipent des formations des équipes de Data Science sur l’IA de confiance
- **36%** ont défini une charte « IA de confiance »

Les entreprises en phase d'approfondissement de la future réglementation d'IA de confiance de la Commission Européenne

Plus de la moitié des participants estiment ne pas être suffisamment informés sur la réglementation de l'IA.

Estimez-vous être suffisamment informés sur la réglementation de l'IA de confiance ?



La plupart des participants se reposent sur des sources d'information internes de l'entreprise qui se basent sur le travail des entités en charge des réglementations, des partenariats externes ou des projets data. Certaines entreprises ont même créé des équipes dédiées pour faire la veille sur les sujets d'IA.

Par contre, la plupart des participants reconnaissent que l'information ne circule pas de la même façon au sein de l'entreprise et qu'il est surtout nécessaire de sensibiliser beaucoup plus les métiers sur l'IA et les impacts potentiels de la nouvelle réglementation.

> Je considère être suffisamment informée sur la réglementation, par contre ce qui manque ce sont des précisions sur sa mise en œuvre opérationnelle au sein de nos organisations pour passer de la théorie à la pratique. Nous devons travailler en interne avec la DPO et en associant nos partenaires, pour décrypter les textes et construire une démarche accompagnant le changement et engageant les différents acteurs du groupe. Ceci se traduit notamment par des actions d'acculturation, de mise en place d'outils de scoring de risque ou encore d'instances de gouvernance de projets.

Chafika Chettaoui,
Groupe Chief Data Officer, Suez

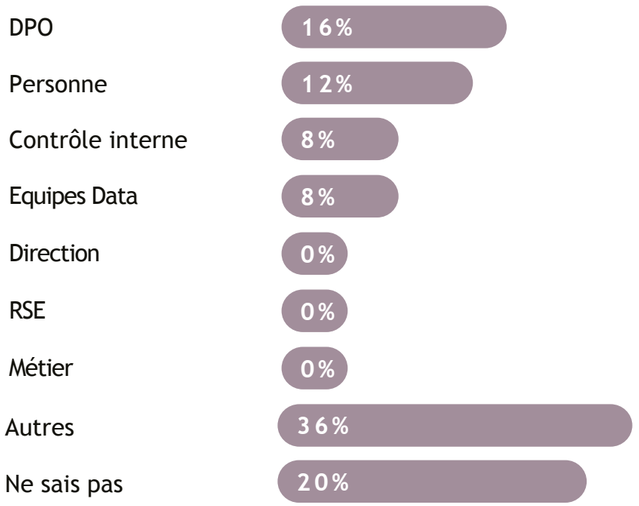
> Une Task Force IA a été créée en janvier 2017. L'un des 6 groupes de travail qui composent cette Task Force réalise une veille active sur les réglementations et contribue également aux consultations ou livres blancs sur le sujet de l'IA de confiance notamment.

Aude Vinzerich,
Directrice du programme IA, EDF

Des organisations peu matures ou non-définies au sein des entreprises pour répondre aux futures exigences d'IA de confiance

Nous avons constaté des réponses très hétérogènes à la question de l'entité en charge du suivi et du pilotage de la réglementation de l'IA au sein de l'entreprise.

Entité en charge de la réglementation de l'IA



De nombreux acteurs internes sont impliqués dans les initiatives d'IA de confiance soit pour des raisons techniques et scientifiques, soit pour des raisons réglementaires. Pour presque la moitié des participants, la responsabilité est partagée par plusieurs directions, par exemple, entre :

- Direction Juridique et Direction des Affaires Européens
- Direction Data, Direction Contrôle interne et Direction Juridique
- Direction Data et Direction Juridique

Pour certains, le Data Protection Officer (DPO) est en charge ou si cela n'est pas encore le cas, devrait être en charge du suivi et du pilotage de cette nouvelle réglementation car cela représente la suite logique du règlement général sur la protection des données (RGPD). Selon les interviewés, la protection et la confidentialité des données, de manière générale, fait partie de l'IA de confiance.

D'autres participants considèrent que ce rôle devrait être porté par l'entité data et son Chief Data Officer ou son Responsable Data Science car :

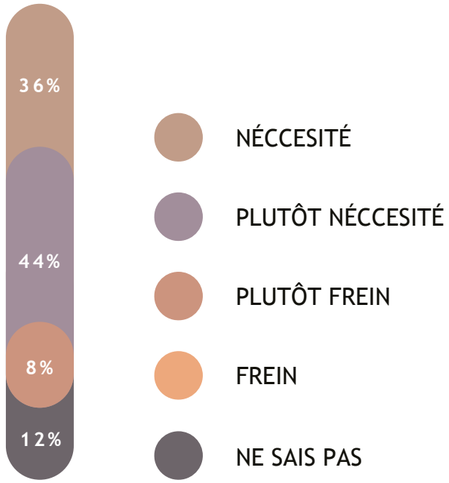
> Certaines initiatives comme, par exemple, l'intelligibilité et la robustesse des modèles, ou le traitement des biais sont déjà lancées par le Datalab.
>
> **Head of Datalab, Banque**

> L'inventaire des modèles au sein de l'entreprise ainsi que le processus de validation de ces derniers est un enjeu clé sur lequel nous travaillons. L'audit des modèles est d'ailleurs un des sujets travaillés et notamment en matière d'expertise requise et d'indépendance des équipes.
>
> **Maxime Havez,**
> **Responsable Département Innovation & IA, Crédit Mutuel Arkea.**

La future réglementation d'IA de confiance considérée plutôt comme une nécessité

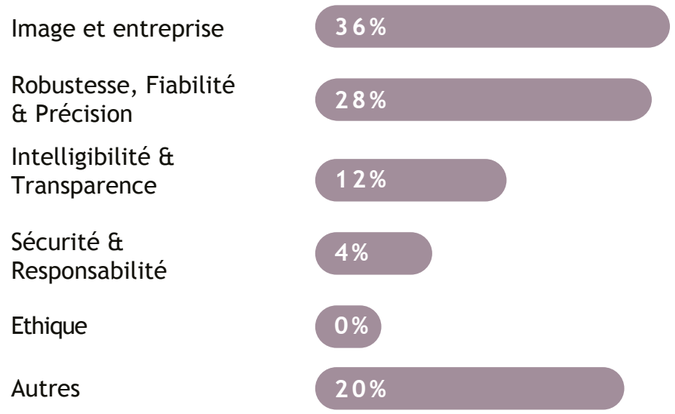
Malgré certaines craintes exprimées par les participants comme la lourdeur du formalisme ou le manque de consignes de mise en œuvre, la plus grande partie des participants pense que cette réglementation nécessaire.

Percevez-vous ce projet de réglementation de l'IA comme une nécessité ou un frein ?



Un tiers des participants considèrent les mesures de l'IA de confiance comme un atout pour son image vis-à-vis des parties prenantes internes et externes de l'entreprise. Les participants pourront envisager d'utiliser des labels pour renforcer la communication mais il est nécessaire de standardiser la démarche des organismes en charge.

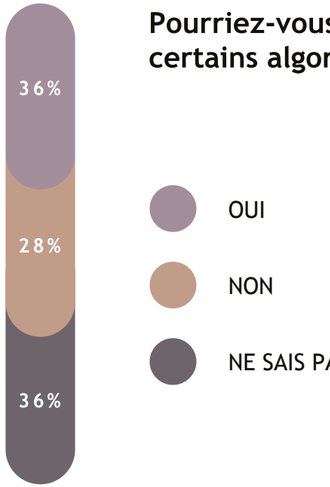
Attraits de l'IA de confiance



L’utilisation de labels d’IA de confiance considérée comme prématurée

Concernant les labels d’IA de confiance, les avis des participants sont plutôt partagés. Un tiers des participants envisagent de labéliser certains algorithmes, un tiers sont plutôt contre et un tiers n’ont pas encore d’avis sur le sujet.

Pourriez-vous envisager de « labelliser » certains algorithmes ?



> Nous pensons qu’une labélisation officielle permettrait d’avoir une meilleure appréciation des ambitions et des réalisations en matière d’IA de confiance.
>
> **Cynthia Traoré,**
> Responsable Data Science et Mireille Deshaye, DPO Swiss Life

> Nous sommes intéressés pour utiliser des labels mais il est trop tôt pour faire le choix. Pour le moment, nous travaillons plutôt sur une charte interne.
>
> Responsable Data Analytics, Assurance

Des impacts financiers importants pour la mise en conformité d’une IA de confiance

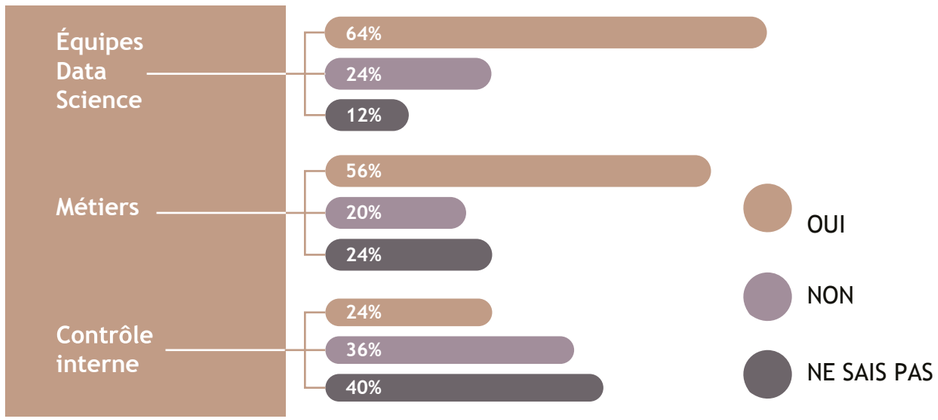
Trois quarts des participants considèrent que la mise en conformité d’une IA de confiance réglementaire aura des impacts financiers. Ceux qui pensent qu’il y aura aucun impact, exercent principalement dans les secteurs non-considérés comme “à haut risque”.

Craintes sur le coût de la mise en conformité

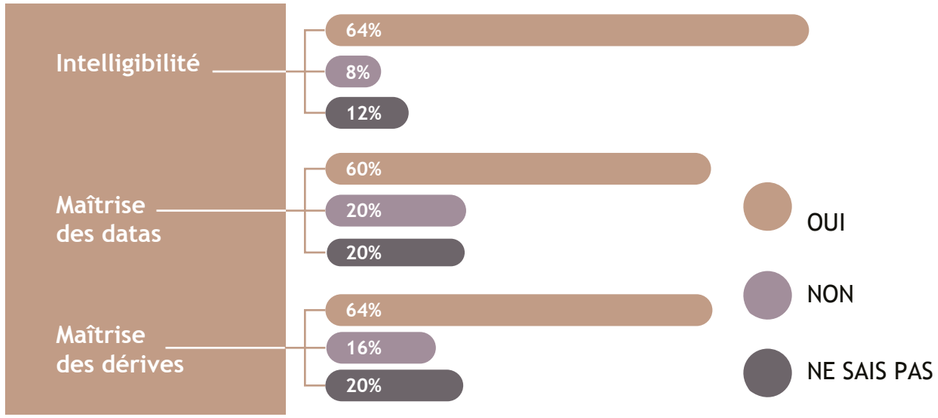
Des formations IA mises en place mais couvrant peu l'IA de confiance

Afin de répondre à la vision et aux ambitions IA de l'entreprise, des formations pour les équipes techniques et des activités d'acculturation pour les métiers sont mises en place, néanmoins très peu traitent les thématiques d'IA de confiance directement comme par exemple les biais des données et des modèles, l'éthique, etc.non-considéré comme "à haut risque".

Anticipation du sujet de la formation et de la sensibilisation des équipes



Mise en place de contrôle lié aux exigences suivantes



La plus grande partie des entreprises ont déjà mis en place des contrôles liés à l'intelligibilité des modèles et plus de la moitié - sur la maîtrise des dérives et des biais.

"La sensibilisation et l'acculturation sur l'IA éthique est essentielle pour les Data Scientists et pour les métiers qui utilisent l'IA mais également pour l'ensemble des collaborateurs."

Cynthia Traoré,
Responsable Data Science, Swiss Life

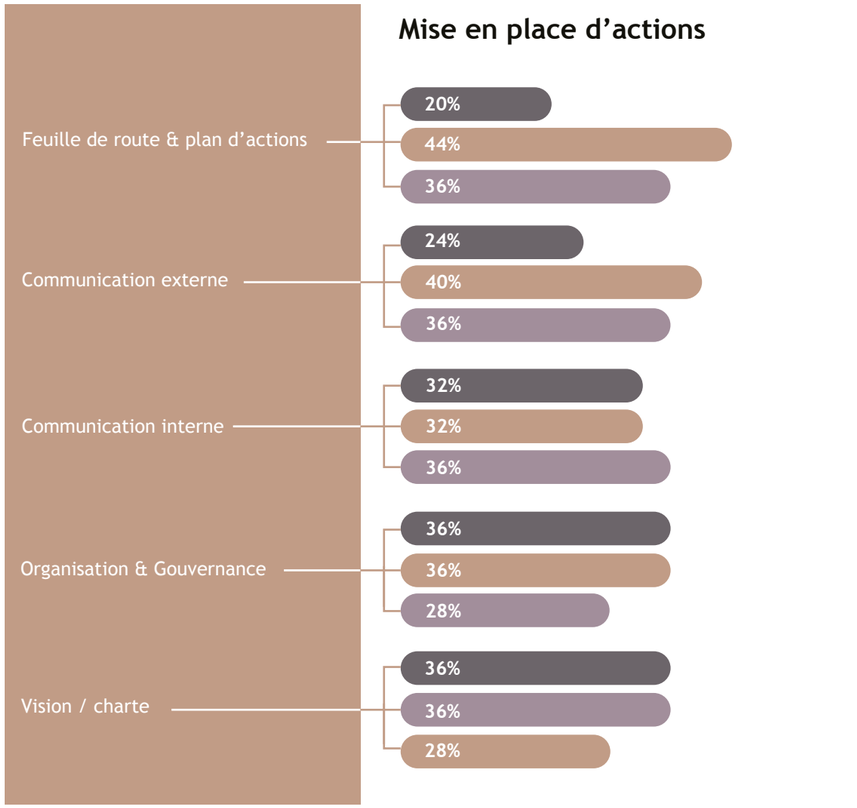
Mireille Deshaye,
DPO, Swiss Life

"L'interprétabilité des modèles est clé. Il faut éviter les effets black box. Nous utilisons SHAP et LIME pour des cas d'usage de churn des collaborateurs."

Architecte, Banque

Peu de mesures organisationnelles et stratégiques mises en place pour répondre aux futures exigences d'IA de confiance

En plus des moyens techniques, quelques mesures organisationnelles et stratégiques ont été mises en place pour répondre aux enjeux d'une IA de confiance comme la charte interne, la gouvernance des équipes ou la communication interne.



Nous constatons que le principal focus des entreprises pour le moment est sur l'industrialisation des solutions IA et leur explicabilité vis-à-vis des utilisateurs finaux. Tous les autres facteurs impactant l'IA de confiance, seront traités avec la mise en œuvre de la nouvelle réglementation de la Commission Européenne. Des nouveaux acteurs seront intégrés dans les équipes IA avec des rôles et responsabilités dédiés au contrôle de conformité. A travers les nouveaux processus et guidelines, l'indépendance entre les équipes créatrices des solutions et celles qui les auditent, sera garantie. Certains participants considèrent cette nouvelle réglementation comme une opportunité de mettre en place des moyens de communication afin de rendre les solutions plus compréhensibles et fiables.

Les mesures organisationnelles et stratégiques considérées comme prioritaires pour 2021

Concernant les priorités 2021 en lien avec l'IA de confiance, les entreprises se focaliseront, de manière générale, sur les moyens organisationnels et stratégiques, et plus précisément :

- Définition de la vision / charte IA de confiance
- Définition de guidelines et processus
- Définition d'une feuille de route IA de confiance
- Coordination des parties prenantes

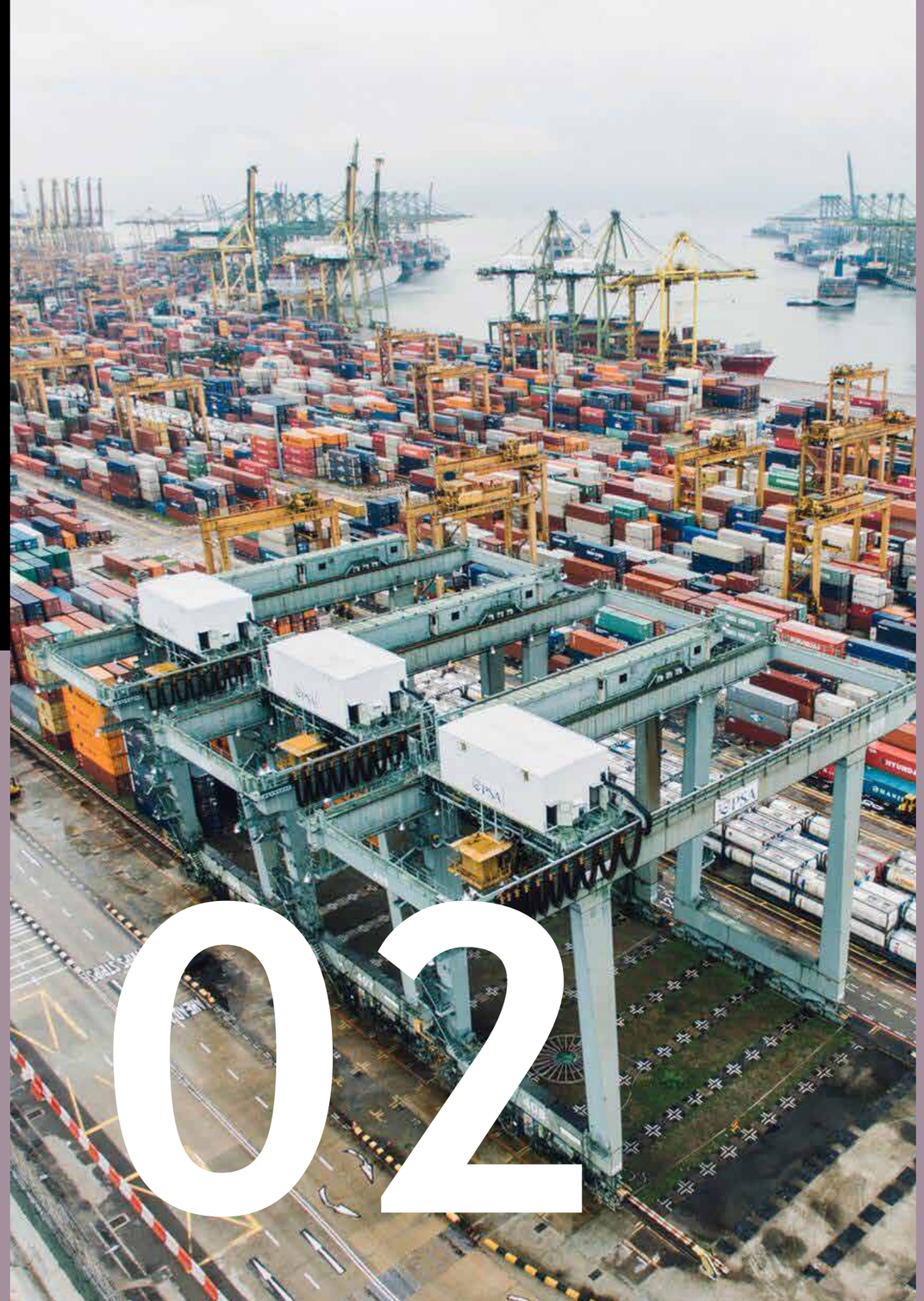
Priorités 2021 IA de confiance

Conclusion de l'enquête

Sans attendre la nouvelle réglementation européenne d'IA de confiance prévue pour 2021, une grande partie des entreprises ont déjà mis en place des moyens techniques pour le suivi des modèles d'IA en production. Les compétences techniques pour assurer la robustesse et l'intelligibilité des algorithmes, sont portées par les équipes de Data Science au sein des entités Data et/ou métier. Des formations ont été mises en place pour assurer la montée en compétence sur l'intelligibilité et la maîtrise des biais dans les modèles et les données.

Néanmoins, ces pratiques couvrent juste une partie de la réglementation européenne en vigueur. La veille sur l'aspect réglementaire et scientifique ainsi que la mise en place des équipes et des processus pluridisciplinaires restent à traiter en priorité afin de pouvoir répondre à l'ensemble des exigences.

Concernant les entreprises qui ne feront pas partie des secteurs ou ne traitent pas des cas d'usage considérés « à haut risque », la vision de l'IA de confiance et la charte éthique restent toujours d'actualité afin de mieux engager non seulement les équipes techniques mais aussi les équipes métier dans une démarche d'IA transparente et éthique. Ces entreprises se focalisent en priorité sur l'image d'IA de confiance auprès de leurs collaborateurs ainsi que de leurs clients et partenaires.



02

Un investissement à valoriser

Notre enquête a révélé un certain nombre de craintes identifiées par les entreprises exploitant la donnée. Ces craintes ne sont pas à négliger mais ne doivent pas éclipser les opportunités liées à une IA plus responsable, notamment concernant leur image de marque et la pérennité des gains liés à la valorisation des données.

Des craintes relatives à de nouvelles exigences

Le cadre juridique actuel posé par l'apport des réglementations de l'UE sur l'IA impacte le déploiement des projets data et l'arrivée d'une réglementation complémentaire fait craindre un alourdissement pouvant décourager les entreprises.

Dans le cadre de la consultation publique sur le livre blanc de la Commission Européenne, les entreprises et les industries ont eu tendance à approuver dans l'ensemble la proposition de la Commission de réglementer des applications d'IA risquées. Pour autant, des opinions contraires ont souligné que de nouvelles exigences ne sont pas nécessaires ou doivent être proportionnées afin de limiter le poids de la mise en conformité et des charges administratives sur les organisations. Des craintes de ce type ont été également soulevées lors de notre enquête.

En réalité, le cadre juridique actuel couvre déjà plusieurs aspects attendus de la future réglementation : le RGPD, en vigueur depuis mai 2018 comporte des exigences sur les prises de décision fondées sur un algorithme, telles que l'intervention humaine dans le cadre de décisions entraînant des conséquences cruciales pour une personne, ou encore le droit pour les personnes d'obtenir des informations sur la logique sous-jacente du fonctionnement de l'algorithme.

A ces directives globales s'ajoutent des textes spécifiques aux secteurs d'activité sensibles et régulés. Pour le secteur bancaire par exemple, la norme Bâle II (2007) en vigueur pour la zone Euro encadre le développement des modèles internes de calcul de risque. Les équipes de modélisation ont des obligations précises incluant la qualité de données en production, le suivi dans le temps du comportement du modèle et la justification des étapes de modélisation. Ces exigences sont proches de celles décrites dans les projets de réglementations actuelles. Les secteurs sensibles ayant un cadre normatif d'ores et déjà en vigueur bénéficieront ainsi des acquis de l'expérience.

Une seconde préoccupation porte sur le manque de consignes claires pour faciliter une mise en conformité concrète et réactive. Pour parer à cela, notre enquête a confirmé le besoin des entreprises d'anticiper les exigences futures pour pouvoir s'engager dès à présent dans la mise en œuvre d'une IA de confiance. Ainsi, les directions ont à spécifier les principes éthiques cadrant la mise en œuvre de leur stratégie afin que les métiers et les équipes techniques puissent les traduire en règles et modalités d'applications. Nous développons une proposition de démarche dans le chapitre 3 de ce livre blanc.

Au niveau européen, des inquiétudes émergent également quant à une perte de l'attractivité du marché européen et l'impact négatif sur l'innovation en Europe, les régulations étant moins contraignantes dans le reste du monde. Pour autant, au-delà d'une exigence visant à protéger les utilisateurs, cette régulation pour une IA de confiance est justifiée par son opportunité de construire un avantage concurrentiel pour l'UE dans la compétition internationale : l'Europe devrait pouvoir développer un avantage compétitif tactique en construisant une troisième voie avec une IA éthique. En effet, l'IA fondée sur des valeurs européennes est un différenciateur sur le marché mondial et un moyen de faire valoir ces valeurs au-delà des frontières de l'UE (comme c'est le cas avec des initiatives telles que le Global Partnership on Artificial Intelligence). Néanmoins, soutenir les entreprises dans la mise en place de leur IA de confiance en évitant l'écueil de brider l'innovation représente un challenge pour la Commission Européenne qui devra accompagner plus que contrôler.

Enfin, d'autres craintes existent également quant à un coût augmenté d'utilisation des IA si des tâches supplémentaires sont à prévoir, et à un risque opérationnel étendu si de fortes amendes sont imposées en cas de non-respect. Cependant, ces coûts sont à mettre au regard des sources de valeur que nous présentons ci-après.

Des sources de valeur à capter

Une réglementation génère des contraintes mais également des opportunités. Les principales sources de valeur à capter par les entreprises sont la confiance de leurs clients et l'image de leur marque, un retour sur investissement durable des solutions IA grâce à une utilisation éclairée par les acteurs opérationnels, et l'accélération de l'innovation sur la thématique sociétale et environnementale.

Confiance des clients et image de l'entreprise

La mise en œuvre d'une IA de confiance est devenue une condition primordiale à la réussite des entreprises mettant en œuvre des algorithmes. En effet, comme l'IA impacte désormais la vie quotidienne des consommateurs, ces derniers s'attendent à des interactions transparentes, équitables et responsables avec les modèles algorithmiques.

C'est pour cela qu'une IA de confiance permet aux entreprises de rassurer les clients ainsi que d'instaurer une plus grande confiance avec eux. Ceci permet d'accroître la satisfaction des consommateurs et par conséquent, de susciter une plus grande loyauté. La fidélisation des clients représente une valeur significative pour les entreprises et peut être soutenue par la mise en place d'une IA de confiance.

Par ailleurs, rappelons que d'après une étude élaborée par l'institut français d'opinion publique en Décembre 2020, plus de 40% des français ressentent de l'inquiétude vis-à-vis de l'IA. Parmi les raisons de ce manque de confiance, l'étude cite l'utilisation malveillante de l'IA, les risques en termes de sécurité privée et de protection des données, la peur que l'humain perde le contrôle sur l'IA et le manque de transparence des algorithmes. L'implémentation d'une IA de confiance pourrait écarter ces sources d'inquiétudes en garantissant plus de transparence et d'éthique vis-à-vis des clients.

Parmi les exemples de dérives qui affectent l'image d'une entreprise, et par conséquent, la confiance des utilisateurs, nous pouvons citer l'exemple de l'IA mise en place par Amazon pour automatiser la sélection de CV pour des postes ouverts. Les grands titres repris et partagés dans les médias ont surtout cité une IA discriminatoire qui ne sélectionne que des profils d'hommes, en omettant l'explication de la faille, à savoir que les données d'entraînement reposent sur un historique de données, où la plupart des candidatures antérieures étaient des hommes. Le biais est donc présent directement dans le jeu de données (Dastin, 2018). Une démarche d'IA de confiance aurait pu permettre d'identifier les risques de biais en amont du projet et de les traiter avant que les modèles ne soient développés et déployés.

Une utilisation éclairée pour une création de valeur pérenne

L'IA existe depuis de nombreuses années et trouve même ses premières applications concrètes dans de nombreux secteurs (santé, retail, etc.) et fonctions d'entreprise (marketing, finance, approvisionnement, etc.) depuis les années 2000.

Cependant, le mode de fonctionnement des entreprises a-t-il été réellement transformé par son utilisation? Même si de plus en plus d'équipes opérationnelles s'approprient des solutions d'IA, elles restent faiblement adoptées par crainte ou manque de confiance. Deux options s'offrent à eux : rester sceptiques et continuer à exploiter des méthodes non basées sur l'IA, ou se fier progressivement aux solutions que les nouveaux algorithmes proposent.

Nous sommes convaincus que l'interprétabilité des modèles et des algorithmes aura un rôle de catalyseur dans l'adoption accélérée des solutions d'IA par les métiers. Elle atténue l'effet "boîte noire" que la plupart des acteurs non intégrés à la conception de la solution IA, perçoit quand il s'agit d'interpréter les résultats obtenus. Une solution d'IA intelligible permet aux organisations d'assurer un meilleur transfert de connaissance et une meilleure prise de décision.

Les exigences liées à l'IA de confiance permettent d'améliorer la compréhension des algorithmes mais apportent également une amplification de leur robustesse, et donc de la création de valeur associée. Une IA intelligible est plus facile à monitorer et à piloter dans le temps. Elle permet non seulement de détecter des problèmes plus facilement, par exemple, les dérives des données, mais aussi de les résoudre plus rapidement, voire de les anticiper afin de minimiser les impacts financiers, humains ou environnementaux générés par ces dérives.

Une solution d'IA intelligible et robuste dans le temps garantit un retour sur investissement (ROI) durable grâce à son utilisation maîtrisée par la majorité des acteurs.

Innovation sociétale et environnementale

La nouvelle réglementation d'IA de confiance aurait un impact sur l'innovation liée à l'IA non seulement sur l'aspect scientifique, technologique et organisationnel pour les acteurs de l'IA mais aussi sur l'aspect sociétal et environnemental pour ses utilisateurs. En effet, l'un des points mis en avant par la Commission Européenne est que la consommation de ressources et d'énergie d'une IA doit être évaluée et remise en cause à chaque étape de son cycle de vie. Cela va sans doute inciter les entreprises à apporter des innovations à leurs modèles afin de réduire leur impact environnemental et construire des solutions fondées sur une IA durable. Cela concerne essentiellement les algorithmes de Deep Learning qui s'appuient sur des sources de données toujours plus volumineuses et des calculs de plus en plus lourds, provoquant de plus en plus d'émissions de C02 engendrées par l'entraînement des dits-modèles. Parmi les solutions innovantes possibles, nous retrouvons l'implémentation d'une IA frugale qui consiste à entraîner les modèles algorithmiques sur des jeux de données de taille réduite. Conscients de l'empreinte énergétique des algorithmes de Deep Learning, de nombreux chercheurs planchent sur cette solution. En outre, comme le témoigne le chapitre 5, plusieurs travaux scientifiques de recherche et de développement sont en cours d'élaboration afin de répondre aux différents points abordés par l'IA de confiance qui sont le traitement des biais, la robustesse des modèles, et l'intelligibilité.

Chartes, Labels et Certifications d’IA de confiance

Étant donné l’importance que prend l’IA au sein des entreprises et l’intérêt croissant pour le sujet d’IA de confiance, de plus en plus d’organismes proposent des labels, chartes et certifications avec l’objectif de pouvoir évaluer le caractère « responsable » des solutions intégrant une IA.

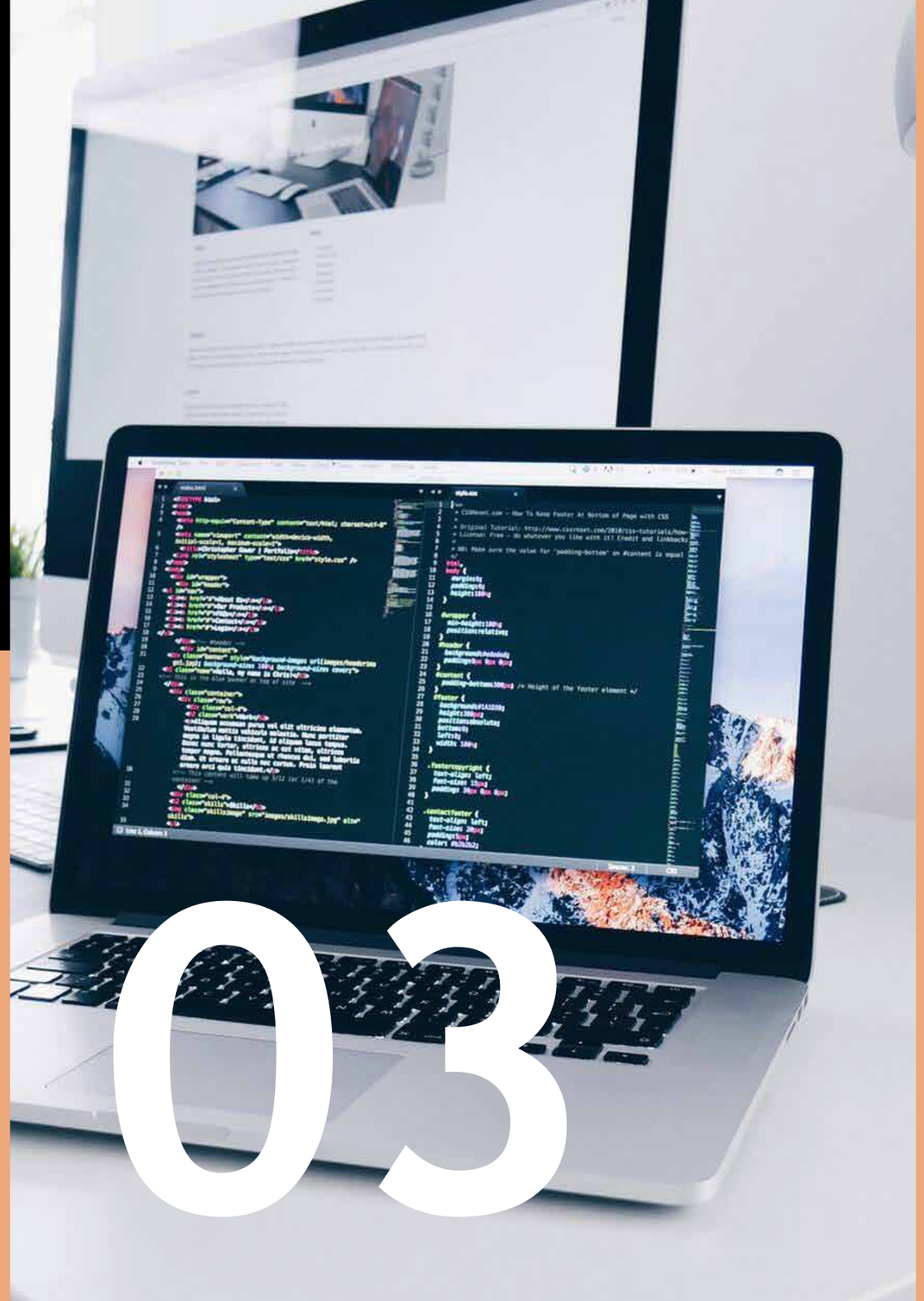
Ces reconnaissances doivent permettre aux entreprises de communiquer sur leur démarche éthique en garantissant à leurs clients la fiabilité de leurs systèmes d’IA.

Différence entre charte, label et certification :

• Une charte est le résultat d’une démarche volontaire impliquant un engagement moral de l’entreprise. Il s’agit en général d’une liste d’actions non soumise au contrôle d’un tiers.
• Un label peut être défini comme un ensemble d’exigences auxquelles les entreprises doivent répondre. Celui-ci peut provenir d’un organisme public ou privé permettant à une entreprise de se distinguer et de communiquer. Par contre, un label est beaucoup moins encadré qu'une certification.
• Une certification est une procédure par laquelle un organisme agréé et extérieur à l’entreprise garantit que celle-ci répond aux exigences d’une norme reconnue.

Les chartes et labels existants se différencient par les secteurs d’activité auxquels ils s’appliquent et les thèmes couverts. Certains se spécialisent sur des enjeux spécifiques tels que environnementaux, de non-discrimination ou même sur des problématiques plus précises tel que l’égalité professionnelle.

Notre conviction est qu’un cadrage normatif, en stimulant l’élévation des attentes et exigences des citoyens, parviendra à imposer progressivement de nouveaux standards. Pour autant, face au foisonnement des labels, un cadrage par la Commission Européenne est souhaitable pour permettre aux citoyens de vérifier facilement la conformité des produits et services à des critères de référence objectifs et normalisés à l'échelle de l'UE.



03

Engager la mise en œuvre et s’inscrire dans la vision et les exigences futures

Google a récemment nommé Marian Croak, PhD, en tant que nouvelle responsable du centre d'expertise "Responsible AI Research and Engineering" selon qui "...le domaine de l'IA responsable et éthique est très nouveau. Ces cinq dernières années, la plupart des institutions n'ont développé que des principes, et ce sont des principes abstraits de très haut niveau. Il y a beaucoup de dissensions, beaucoup de conflits pour essayer de standardiser les définitions normatives de ces principes. Quelle définition de l'équité ou de la sécurité allons-nous utiliser ? Il y a actuellement beaucoup de conflits sur le terrain, et cela peut parfois être polarisant..." (Croak, 2021)

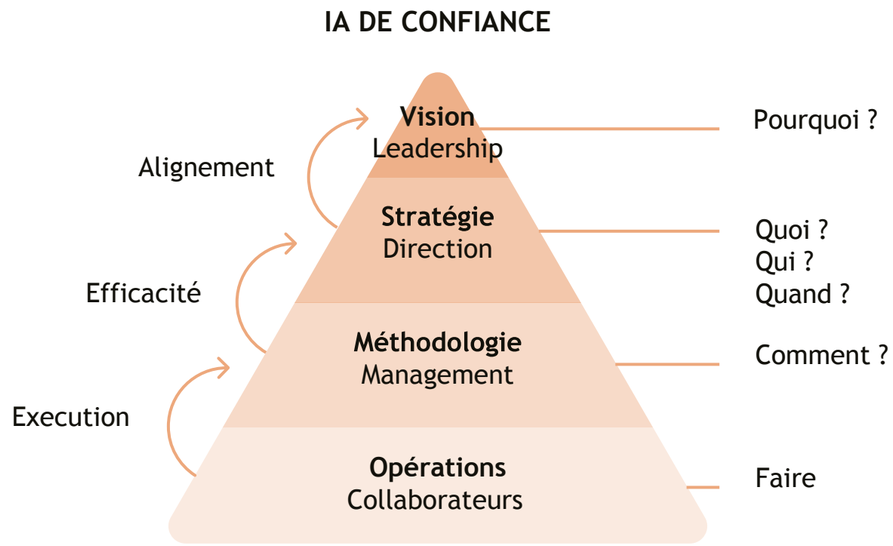
Comment les entreprises peuvent-elles alors s'appuyer sur ces principes abstraits très haut niveau et mettre en œuvre une IA de confiance ?

Les entreprises deviennent de plus en plus data-driven : elles expriment de fortes ambitions de valorisation des données générées par les activités de leurs différents métiers. Nous observons l'émergence de nouvelles structures et de rôles dédiés au lancement et au déploiement des initiatives IA. Selon la maturité des entreprises, les directions Data peuvent être positionnées soit au niveau central, soit en silos avec des rôles directement intégrés dans les entités métiers, soit sous un modèle hybride comme présenté dans notre baromètre des organisations Data 2020.

Nous présentons ici quelques orientations à destination de ces organisations afin de concevoir et déployer une IA de confiance par la prise en compte des 7 exigences aux différents niveaux de la structure de l'entreprise. Ces 7 exigences ne sont pas encore entrées dans le cadre législatif mais vont servir de fondation pour les prochains règlements européens sur l'IA. S'en servir comme base de référence pour une IA de confiance en entreprise permet donc d'anticiper ces règlements.

Les approches proposées ici peuvent être envisagées isolément ou ensemble selon la situation de l'entreprise. D'autres approches ont leur raison d'être, celles présentées ici se veulent dans la lignée des recommandations du groupe d'expert AI HLEG.

Figure 5 : Déclinaison de l'IA de confiance au travers de la pyramide hiérarchique d'une organisation



Cadrer une démarche pour établir une IA de confiance

Spécifier et partager la situation actuelle de l'entreprise au regard de l'IA

Afin de définir le positionnement de l'entreprise vis-à-vis des enjeux d'IA de confiance, impliquer un groupe de travail 'core team' pluridisciplinaire semble pertinent pour établir un état des lieux partagé et définir la marche à suivre. A titre de référence, ce groupe de travail peut regrouper les directeurs IA/data et SI mais aussi des directeurs d'autres entités, notamment Risque et conformité, Juridique, RSE, Communication, Entités métiers et divisions propriétaires de solutions d'IA.

L'état des lieux peut être réalisé en évaluant plusieurs axes d'analyses, notamment :

Sensibilisation aux enjeux et risques de l'IA : Quel niveau d'acculturation et de formation des acteurs concernés par l'IA et plus généralement des collaborateurs ? sont-ils initiés à l'IA, sensibilisés aux risques associés ?

Connaissance de la réglementation : Par quel cadre réglementaire sommes-nous concernés ? Quelles sont les recommandations des organismes de régulation au sujet de l'IA de confiance ?

Maturité de l'environnement : Quelles actions sont entreprises par les entreprises comparables, concurrentes, ou leader dans le domaine de l'IA ? Où en est la recherche scientifique ?

Existant et initiatives déjà mises en place : Quelles actions déjà réalisées peuvent être pérennisées ? Quelles initiatives transverses, similaires à celles relevant du RGPD par exemple, doivent être lancées ? Quelles directions sont déjà impliquées et quels sont leurs rôles et responsabilités ?. Exemples d'acteurs internes et rôle associé :

- **IA et data :** Coordinateur / Contributeur - son objectif est de réaliser des solutions performantes et robustes ; en charge des méthodes de gouvernance des modèles et des données.

- **Système d'Information :** Contributeur - responsable des systèmes et des infrastructures ; son objectif est de faciliter l'industrialisation et le run des solutions.

- **Risque et Conformité :** Coordinateur / Contrôleur - identifie, évalue et contrôle l'ensemble des risques de manquement aux obligations réglementaires.

- **Juridique :** Coordinateur – pilote, anime et coordonne le conseil juridique interne ; anticipe les risques liés à la propriété intellectuelle et les responsabilités juridiques.

- **RSE :** Coordinateur - responsable du suivi et du pilotage des risques économiques, sociaux, sociétaux, environnementaux et leur impact sur la réputation et l'image de l'entreprise, ainsi que du climat de confiance avec les parties prenantes et de l'engagement des collaborateurs.

- **Relations publiques et communication :** Contributeur – responsable de la communication externe et des relations publiques ; son objectif est de participer activement aux initiatives externes et de communiquer là-dessus dans une démarche de création d'une réputation et d'une image de l'entreprise.

Les autres directions métiers : Contributeur – responsable de l'application et de l'usage des solutions

L'instruction de ces différentes thématiques peut se concrétiser par la formalisation d'une charte éthique relative à l'intelligence artificielle et ses applications au sein de l'entreprise s'appuyant sur l'état des lieux réalisé et explicitant l'ambition de l'organisation sur ces sujets.

Définir les principes d'une IA de confiance et les décliner opérationnellement

Les exigences s'appliquent rarement sans interprétation et priorisation adaptée au contexte de l'entreprise. On conçoit aisément l'importance de l'implication de la direction générale dans cette transcription des exigences et dans l'affirmation de sa volonté. Il est donc primordial que le Top management soit partie prenante de la définition et validation de cette feuille de route

Spécifier les principes d'une IA de confiance cadrant la mise en œuvre de la stratégie permet aux équipes opérationnelles (métiers comme techniques) de les traduire en règles et modalités d'applications. Ces principes ont ensuite vocation à se concrétiser en exigences de référence pour l'entreprise.

Ainsi par exemple, le principe de "Transparence autour du traitement" pourrait être spécifié en deux exigences de référence : "Informer les utilisateurs lors de l'utilisation d'une IA", et "Fournir une fiche explicative pour tout algorithme d'IA produisant une décision".

Au regard de ces principes et exigences de référence, on définira l'organisation pratique :

- Quels sont les rôles clés à créer dans l'organisation pour mettre en place une IA de confiance ?
- Quels sont les instances et les processus à créer qui régiront l'usage de la donnée et de l'IA pour garantir la confiance ?
- Comment définir et suivre les standards de l'IA de confiance ? À quel moment et comment faut-il les implémenter ?
- Quels outils sont à utiliser pour faciliter la mise en place et le pilotage d'une IA de confiance ?

Cette déclinaison de l'ambition peut prendre la forme d'un **référentiel d'exigences**, amené à évoluer selon la réglementation et les progrès des méthodes d'intelligence artificielle.

Orienter l'évolution des pratiques en termes d'IA de confiance

Former et sensibiliser aux enjeux et méthodes

L'appropriation des enjeux et du fonctionnement de l'IA par les équipes impliquées est un prérequis évidemment nécessaire à toute transformation pérenne de son utilisation.

Pour y parvenir, la mise en place de formations spécifiquement adaptées aux enjeux de l'IA de confiance (Intelligibilité, robustesse, biais, …) ainsi qu'aux différentes équipes (selon le rôle qu'elles auront à jouer : développement de solutions, mise en œuvre, audit etc.) garantit un accompagnement progressif de la maturité de l'entreprise dans le domaine.

Ce **programme de formation** devra se concentrer sur les acteurs ayant un rôle directement lié avec les systèmes IA de l'entreprise. Les concepts d'IA de confiance pourront néanmoins être évoqués également au sein du cursus d'acculturation du reste des collaborateurs, le sujet restant éminemment sociétal.

Identifier et qualifier les risques pour prioriser la mise en œuvre des exigences

La diversité des risques liés à l'IA requiert la contribution d'acteurs variés pour les identifier et les qualifier correctement avec un spectre d'analyse le plus large possible. Selon les compétences des uns et des autres, l'accent sera mis sur telle ou telle des 7 exigences d'une IA de confiance.

Cette identification et qualification des risques peut être réalisée pour chaque processus utilisant l'IA sur la base des 7 exigences décrites par l'UE complétées du référentiel d'exigences de l'entreprise et devra par ailleurs être mise à jour à chaque jalon clé des projets impactant ce processus.

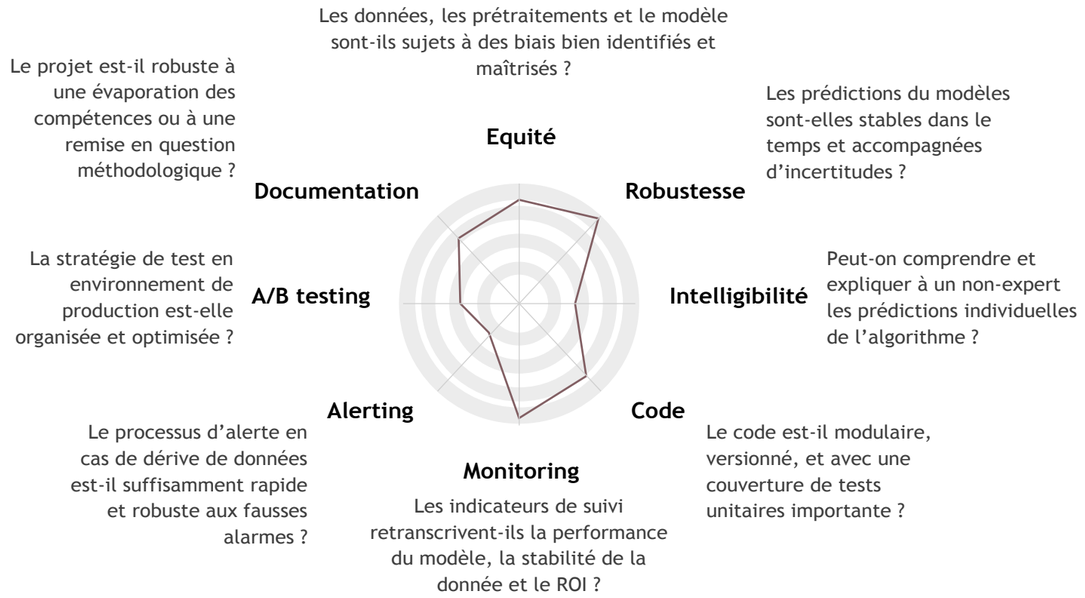
La criticité des risques s'évalue classiquement en fonction de la probabilité d'occurrence, la gravité et la probabilité de non-détection. L'évaluation de la gravité inclut des dimensions éthiques particulières ; ainsi pour ce qui concerne les décisions à portée individuelle (scoring de crédit, scoring des employés etc.) une attention est portée sur le caractère de dépendance du sujet soumis à la décision et sur le dommage potentiel individuel voire sociétal.

Mettre en place une pratique d'audit des modèles

Les contrôles internes des solutions d'IA en référence aux exigences internes et réglementaires se feront autant que possible avec des équipes de vérification différentes de celles de réalisation des solutions (peer review) ou le cas échéant par une équipe transverse dédiée au contrôle de l'ensemble des solutions d'IA.

Schématiquement, l'audit d'un modèle débute par la définition d'une méthodologie reposant sur un ensemble de points de contrôle permettant de couvrir une vue complète des enjeux (métier et technique) et de définir des recommandations et des leviers à activer pour résoudre les faiblesses du système d'IA.

Figure 6 : Exemples de critères d'évaluation d'un modèle IA
Comment auditer une IA ?



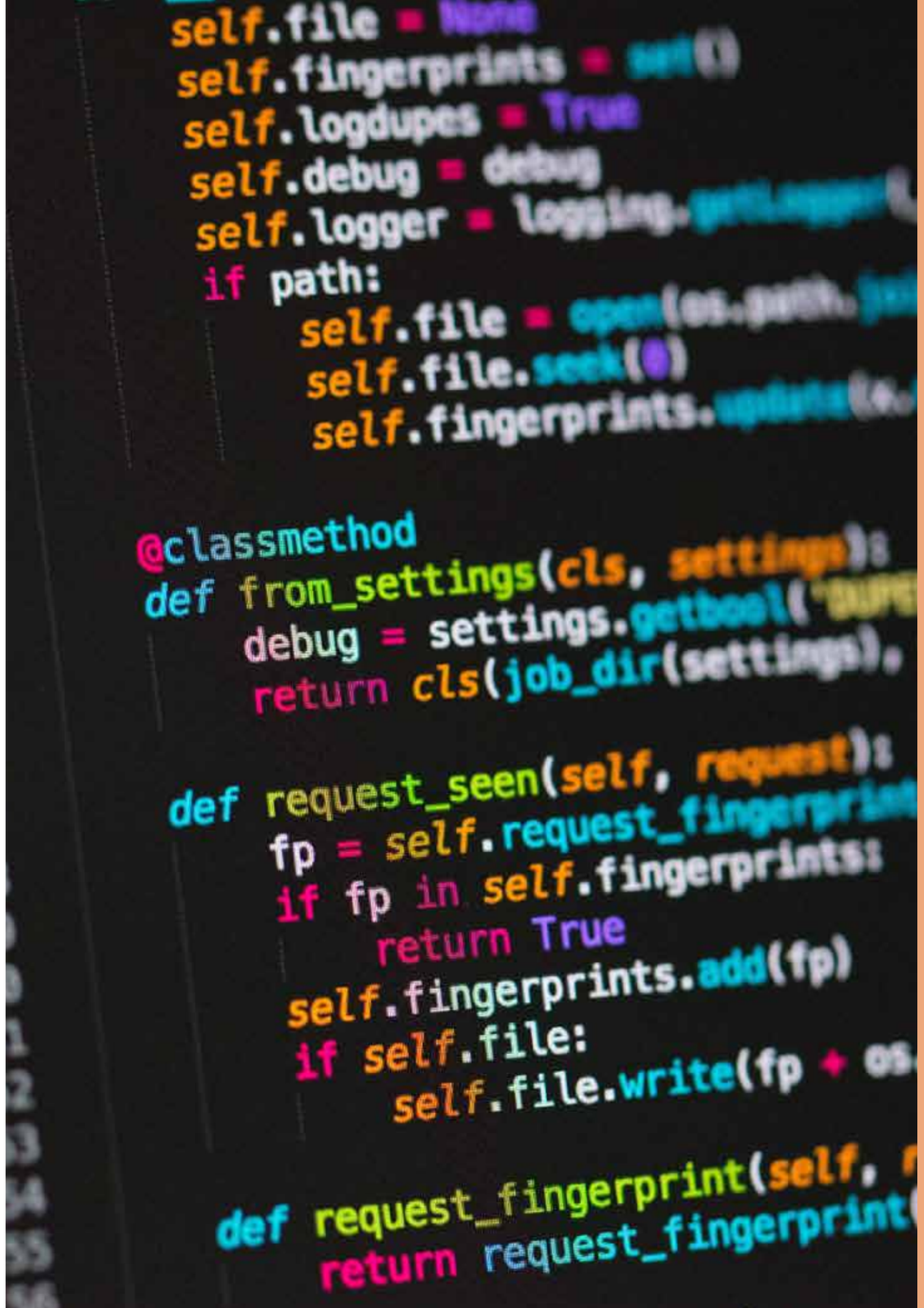
Cette méthodologie d'audit sera utilisée par les équipes de contrôle selon un plan de contrôle permettant une supervision pendant la conception du système mais également après la mise en production afin de contrôler le cycle de vie des modèles.

Établir et entretenir la confiance

Comme nous venons de le voir, il n'est pas possible de donner confiance dans un système basé sur l'IA sans garantir sa maintenance et son amélioration continue en fonction des évolutions à la fois du processus d'entreprise dans lequel il s'intègre et de son environnement.

Les évolutions du modèle post-production relèvent des retours d'expérience successifs (à la fois sur la méthode et les résultats), des évolutions dans le fonctionnement, l'organisation et la stratégie de l'entreprise et des évolutions des attentes et exigences externes (réglementaires ou commerciales par exemple). Ces améliorations portent également sur l'intégration de nouvelles technologies plus fiables (nouveaux modèles, ajout de surcouches d'intelligibilité, …), à l'état de l'art en matière d'IA de confiance.

Afin d'entretenir la confiance, l'information et la communication proactive auprès des utilisateurs et bénéficiaires sont primordiales afin de leur permettre d'apprécier et de valoriser les engagements de l'entreprise dans l'IA de confiance..



04

Cas d’usage illustrés

Afin d'illustrer la mise en œuvre d'une IA de confiance on trouvera ci-après 3 cas d'usages présentant quelques bonnes pratiques de conception.

Figure 7 : Cas d'usage illustrés d'IA de confiance

Cas d'usage	Octroi de crédit	Scoring client	Maintenance prédictive
Problématique	Maîtriser les biais pour une utilisation éthique de l'IA.	Comprendre le fonctionnement d'un algorithme avec l'intelligibilité des modèles.	Assurer la robustesse et le cycle de vie d'un modèle afin d'être prêt pour sa mise en production.

L'éthique, comment développer une solution sans biais

Il est nécessaire de pouvoir contrôler la capacité des modèles à ne pas produire ou reproduire des décisions qui seraient en contradiction avec notre éthique sachant que des biais peuvent surgir à n'importe quel moment de la chaîne de production. Prendre en considération cette problématique est capital pour développer une solution qui soit non-discriminatoire, équitable et qui ait un impact sociétal positif.

Introduire le développement d'outils éthiques sensibilise l'ensemble de l'entreprise à prendre conscience de ses impacts sociétaux et encourage leur utilisation. Au-delà des enjeux sociétaux, investir dans l'éthique s'avère rentable : en terme d'image de marque, les clients auront confiance dans le fait d'être traités de façon non-discriminatoire, juste et honnête. Et en se mettant à l'abri de mauvaises décisions pouvant avoir à terme des conséquences juridiques, l'entreprise évite d'hypothéquer son avenir.

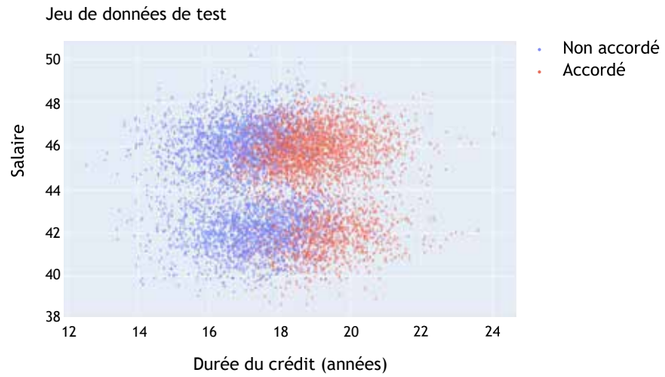
Cas d'usage : Octroi de crédit chez MaBanquePréférée

MaBanquePréférée est une banque qui délivre à ses clients des crédits immobiliers. Pour pouvoir délivrer ces crédits, la solvabilité du client est estimée grâce à un modèle simple de régression logistique. La banque dispose de plusieurs années d'historique de données client. S'appuyant sur l'ambition réglementaire de la Commission Européenne, MaBanquePréférée souhaite vérifier que cet algorithme suive des principes éthiques afin d'éviter de discriminer certaines populations.

Étape 1 : Anticiper les biais et identifier les populations à risque

Tout d'abord, le management et les conseillers travaillent de concert pour identifier les populations potentiellement sujettes à un traitement non-équitable par le modèle. L'objectif est ensuite, pour les équipes techniques, de cibler un potentiel biais et de vérifier s'il est discriminant ou non. Ici, on vérifie que l'algorithme n'est pas discriminant en fonction du genre. Pour cela, les analystes observent les données utilisées dans le modèle (le salaire et la durée du crédit) ainsi que les décisions induites (crédit accordé ou non).

Figure 8 : Cas d'usage « Octroi de crédit » : jeu de données de test



Au-delà de la frontière de décision de l'algorithme (rouge/bleu), une nette frontière est observable entre le nuage de points du haut et celui du bas. Après analyse, le nuage de points du bas correspond aux données des femmes et celui du haut à celles des hommes. Ici, la variable du salaire constitue une variable tampon, on peut différencier une femme d'un homme avec sa donnée salariale. On peut alors distinguer que les femmes sont défavorisées par deux biais :
Biais historique : la différence salariale entre les femmes et les hommes
Biais de représentation : le jeu de données d'entraînement a moins d'observations de femmes que d'hommes (le nuage de point du bas compte moins de points)

Étape 2 : Prendre conscience de l'impact de ces biais en quantifiant les effets discriminants sur les populations sensibles identifiées

Afin de mesurer l'impact de ces biais sur l'algorithme de crédit, l'équipe de Data Scientists calcule le DI (disparate impact). Un modèle non-discriminant aura un DI de 1. Cela veut dire ici que la probabilité d'octroi de crédit est égale quel que soit le genre du demandeur.

$$DI = \frac{P(Y = 1/unprivileged)}{P(Y = 1/unprivileged)}$$

La direction a décidé que le DI de l'algorithme doit être compris entre 0,8 et 1,25 pour valider sa mise en production. Grâce à cette métrique, ils peuvent ainsi maîtriser l'équilibre équité / performance.

Étape 3 : Appliquer une méthode de correction pour corriger ces biais

Afin de corriger ces biais, différentes méthodes de correction existent. Celles-ci se différencient en fonction du moment où cette correction est apportée :

- Pré-processing : Nous venons modifier les données d'entrée du modèle
- In-Processing : Le biais d'équité est minimisé pendant l'entraînement
- Post-Processing : Une fois le modèle entraîné, nous venons transformer les scores de prédictions

Les Data Scientists comparent ces 3 approches. Leur première méthode consiste à exclure la variable genre des variables du modèle ; leur deuxième méthode consiste à appliquer plus de poids aux femmes ayant reçu un crédit ainsi qu'aux hommes n'en n'ayant pas reçu et leur dernière méthode vient ajouter une contrainte éthique dans l'entraînement du modèle. On rassemble ici les résultats du DI et de la métrique de performance (l'accuracy dans notre cas) :

Figure 9 : Cas d'usage « Octroi de crédit » :
Résultats du disparate impact et de l'accuracy

Stratégie	Effet Disparate	Accuracy
Modèle initial	0.547	0.853
Sans variable genre	0.749	0.82
Repondération	0.961	0.812
Contrainte d'entraînement	0.932	0.821

Tout est question de balance d'équité et de performance. Seules les deux dernières méthodes ont un DI acceptable pour l'entreprise. La performance est moindre, et c'est normal. Ici aussi, nous sommes en présence d'un biais, celui d'évaluation. Les jeux de données d'entraînement et de test ont été construits à partir de la même population, nous retrouvons donc les mêmes biais dans les deux jeux de données. La métrique de performance traduit donc la capacité de l'algorithme à reproduire les biais passés. Cette métrique mesure finalement la discrimination globale entre deux groupes. Pour résoudre complètement la problématique discriminatoire, il faut également utiliser l'intelligibilité pour une étude locale du problème.

Intelligibilité et transparence : comprendre de A à Z le fonctionnement de l'algorithme

Quand on parle d'IA de confiance, on pense évidemment à l'intelligibilité. Éviter l'effet « boîte noire » d'un algorithme est la priorité numéro un pour établir la confiance dans le modèle. Le Data Scientist autant que le responsable de projet, l'utilisateur final ou le responsable juridique / DPO se posent chacun des questions sur le fonctionnement de l'algorithme. L'adoption de méthode d'intelligibilité permet alors d'étudier certaines variables sensibles et de repérer les angles morts du modèle afin d'assurer sa fiabilité et in fine de donner confiance à l'utilisateur non technique et de faciliter la prise de décision opérationnelle.

Plus que la compréhension du modèle, il s'agit aussi d'être totalement transparent sur l'ensemble de la chaîne de traitement, des données d'entrée du modèle, aux métriques de performances utilisées jusqu'aux prédictions. L'utilisateur doit avoir en sa possession toutes les clés de compréhension du processus. C'est à l'aide de manuels d'utilisation qu'on apprend à se servir d'outils. Par analogie, tout modèle d'IA doit avoir son manuel à disposition.

Cas d'usage : Scoring client dans le département marketing de WearIt

WearIt est une entreprise dans le retail qui vend un large panel de vêtements au grand public. Depuis quelques mois, ses clients traditionnellement fidèles viennent de moins en moins souvent en magasin. L'objectif marketing a été défini : il faut réactiver les clients avec des promotions très attractives. Le budget alloué pour ces promotions est restreint, il faut donc bien choisir les clients qui seront les plus susceptibles de revenir en magasin. Au-delà d'un ciblage en fonction de la probabilité de retour en magasin, la direction marketing souhaite identifier des leviers d'activation en comprenant les décisions du modèle.

Étape 1 : Définir avec les métiers leurs besoins de compréhension et le cadre d'application du modèle

En premier lieu, les équipes techniques doivent se réunir avec les analystes de WearIt afin de définir les champs d'application de l'algorithme ainsi que le niveau d'intelligibilité qui les intéresse. Le responsable de projet et les équipes métiers vont pouvoir alors s'interroger sur leur attendu. Quelles variables sont les plus importantes pour le modèle et comment influencent-elles les prédictions ? Comment peut-on expliquer une prédiction donnée ? Quels éléments ajouter à la prédiction brute pour que la prise de décision soit fiabilisée et efficace au maximum ? Sous quelle forme restituer l'information à l'utilisateur final du modèle ?

L'objectif premier de WearIt est de comprendre pourquoi certains clients ne sont pas prédits comme à haut potentiel. Qu'est-ce qui leur manque pour qu'ils le soient ? Tirer des règles simples pour comprendre la prédiction sur une population spécifique leur permettrait également de prendre des actions ciblées.

Étape 2 : Choisir le bon algorithme qui satisfait à la fois la demande d'intelligibilité et la performance cible

Faire le bon choix d'algorithme, c'est choisir un modèle dont la performance sera acceptée par le métier et dont l'intelligibilité répond à leur demande. Les modèles se distinguent en deux familles :

1. Les modèles intelligibles par nature (la régression linéaire, les modèles additifs généralisés, les arbres de décisions, etc.).
2. Les modèles dits « boîte noire » (les modèles ensemblistes, les modèles de deep learning, etc.) pour lesquels une surcouche d'intelligibilité est nécessaire, et à prendre en compte lors de la mise en production.

Après avoir essayé différentes modélisations, les Data Scientists décident de retenir un modèle fondé sur l'algorithme Skope Rules pour déterminer les clients les plus susceptibles de revenir en magasin. Skope Rules permet de déterminer des règles interprétables réalisant un compromis entre la meilleure précision et une couverture maximale des données d'apprentissage. Par son fonctionnement, Skope Rules permet d'obtenir une interprétabilité équivalente à celle d'un arbre décision tout en mettant à profit le pouvoir de modélisation d'une forêt aléatoire. Les Data Scientists proposent ainsi un modèle performant, dont les règles définissent directement des sous populations auxquelles les métiers associent des leviers d'actions personnalisés.

Afin d'orienter les équipes métier pour réengager les clients non ciblés (c'est à dire les clients les moins susceptibles de revenir en magasin selon le modèle), les Data Scientists proposent également de déterminer quelle(s) variable(s) modifient la prédiction de ces clients. Les explications contrefactuelles nous permettent de répondre à cette question. Cette approche génère, pour un client qui ne serait pas ciblé, des exemples très similaires à cette observation mais dont la prédiction du modèle aurait été positive. Ainsi, pour certains clients non ciblés, les exemples contrefactuels ont montré que l'abonnement à la newsletter d'emails les auraient prédits en positif. Travailler sur la publicité fonctionne mais il convient d'accentuer les efforts sur son adoption.

Étape 3 : Assurer la transparence de la conception algorithmique

Pour avoir pleinement confiance dans un algorithme, sa conception doit être transparente et comprise de tous. En plus d'accompagner l'utilisateur dans sa compréhension du modèle, il est nécessaire au regard des audits de garantir la pertinence de sa conception. Cette transparence suppose une documentation comprenant - à titre indicatif et non exhaustif - les éléments suivants :

- Les sources de données utilisées
- Les étapes de data processing et les filtres sur les données d'entrée
- Les étapes de feature engineering
- Une description du jeu d'entraînement et de test
- Le modèle utilisé et la façon dont il est implémenté
- Les métriques de performance
- La méthode d'évaluation des incertitudes
- Les modèles d'intelligibilité s'il y en a
- Les étapes de post processing des prédictions
- Les usages attendus et les usages proscrits
- Les limitations et recommandations d'usage
- Les systèmes qui accueillent les prédictions

Les attentes de chacun (direction, métier, technique) en termes d'intelligibilité et de transparence seront ainsi satisfaites. Le travail de transparence s'avèrera par ailleurs déterminant lorsque le modèle en production devra être ajusté.

Cycle de vie du modèle : Assurer son suivi afin de rester efficace

L'objectif de tout projet d'IA est d'arriver à cette étape tant attendue : la mise en production. Cette mise en production doit être suivie par une stratégie organisée, planifiée, pour anticiper la gestion du cycle de vie du modèle. Ceci permettra aux équipes techniques d'affirmer en toute tranquillité les performances attendues de l'algorithme, connaître ses limites et assurer un ROI durable.

Organiser une stratégie de cycle de vie du modèle permet d'**assurer un contrôle humain**. La **robustesse** du modèle sera un atout majeur pour rester confiant dans ses prédictions et permettre aux équipes, métiers et techniques d'anticiper son comportement.

Cas d'usage : Maintenance prédictive chez MonVol

MonVol est une entreprise de construction d'avions. Afin d'éviter que ses machines tombent en panne et arrêtent l'ensemble de la chaîne de production, les Data Scientists ont développé un modèle de machine learning permettant de prédire la survenance des pannes à partir de données de capteurs.
Ce modèle sera bientôt mis en production.

Étape 1 : Assurer la robustesse du modèle avant de le mettre en production

Le modèle de détection de pannes est prêt, enfin presque prêt. Nous ne savons pas encore si le modèle est assez robuste pour être envoyé en production. Nous savons qu'un modèle de machine learning est incertain, le modèle parfait n'existe pas. En revanche, ce dont on peut être (presque) sûr, c'est le taux d'erreur de l'algorithme, si sa robustesse est vérifiée.

Dans un premier temps, on doit contrôler le sur-apprentissage temporel du modèle. L'idée est de valider les performances de l'algorithme par rapport à des jeux de données de périodes temporelles différentes à celle du jeu d'entraînement. Des jeux d'entraînements et de tests sont créés par fenêtres glissantes. On mesure la capacité de généralisation de l'algorithme dans le temps.

Il est également important de considérer l'incertitude du modèle sur ses prédictions. Si le modèle prédit qu'une panne va survenir dans 100 jours 2 jours, MonVol peut alors raisonnablement attendre le prochain contrôle technique. En revanche, si le modèle prédit la survenance d'une panne dans 100 jours 90 jours, MonVol opérera un contrôle technique dès à présent. La méthode Jackknife+ permet de calculer cette incertitude et est agnostique du modèle utilisé. Fort d'un fondement mathématique avéré, cette approche de ré-échantillonnage permet d'avoir une estimation de l'incertitude en entraînant des modèles sur des sous-échantillons différents. Être certain de son incertitude, c'est anticiper l'erreur connue du modèle d'IA.

Étape 2 : Définir les métriques de monitoring

On peut maintenant définir les métriques à suivre sur l'ensemble du cycle de vie du modèle. Ces métriques seront monitorées et versionnées afin d'alerter d'une quelconque dérive. On distingue deux types de métriques :

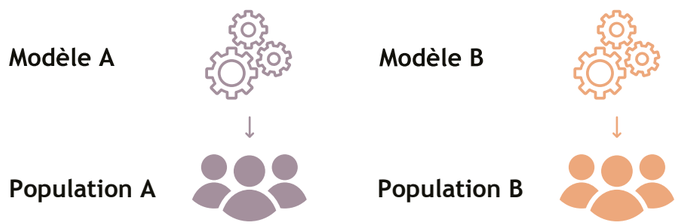
Les métriques métier: les premières métriques à suivre sont celles définies entre les équipes techniques et les métiers qui ont validé les résultats du projet. Ces métriques rassemblent tous les indicateurs ayant permis de valider la mise en production du modèle. Les équipes métier étant sensibles à la pré-détection au plus tôt d'une panne, l'erreur (MAE : mean absolute error) est calculée dans le temps pour mesurer le délai optimum. Le modèle doit fournir une prédiction par jour sur les 25 machines de l'atelier de MonVol. Un indicateur doit également permettre de suivre la livraison effective de ces prédictions.

Les métriques de dérives : les deuxièmes types de métriques sont celles liées à la dérive des données et du modèle qui permettent de comprendre la variabilité des métriques métier. Des tests unitaires permettent de vérifier la qualité des données d'entrée en définissant des seuils avec le métier. La distribution de ces données ainsi que des prédictions doivent rester semblables à celles du jeu d'entraînement. Il faut comparer ces distributions et lever une alerte si celles-ci sont sensiblement différentes. Finalement, la robustesse est à réévaluer constamment pour garder confiance dans les prédictions du modèle.

Étape 3 : Établir une stratégie de réentraînement du modèle

Les métriques de monitoring ayant été définies entre les équipes métier et techniques, il reste à savoir quand il convient de ré-entraîner le modèle. Des seuils de dérives sur les métriques sont définis pour alerter et indiquer qu'un réentraînement est nécessaire. Le temps moyen entre les pannes étant supérieur à l'année, on obtient difficilement de nouvelles observations labellisées. La méthode canary est alors privilégiée.

Figure 10 : Cas d'usage « Maintenance prédictive » : La méthode canary



La méthode Canary consiste à déployer sur une sous population le modèle que l'on souhaite passer en production. Le modèle existant est appliqué à 99% de la population cible et le nouveau modèle est appliqué au pourcentage restant, tiré aléatoirement de la population globale. Le risque généré en déployant ce nouveau modèle est ainsi réduit et sa performance peut être évaluée afin de décider, s'il peut remplacer le modèle précédent.

Après avoir réalisé ces étapes, le modèle part alors en production. Les Data Scientists sont confiants dans ces prédictions et les équipes dirigeantes peuvent mesurer le gain apporté. Le comportement de l'algorithme est sous contrôle.

05

Avancées récentes et perspectives de la recherche

Nous commencerons d'abord par la notion d'éthique, qu'il est possible de caractériser quantitativement par la recherche des biais et de leur correction, puis nous aborderons les différents aspects de la robustesse. Nous adressons ensuite l'intelligibilité pour finir par le cycle de vie du modèle.

Éthique et correction de biais

Comme nous l'avons vu, la correction des biais consiste d'abord à identifier les natures et origines possibles des biais : biais de préparation des données, biais de sélection, de représentation, biais à l'inférence, biais à l'évaluation historique et biais d'évaluation. Le problème de la mitigation du biais nécessite d'être capable d'identifier précisément les populations à risque et les variables sensibles du modèle. La stratégie de correction dépend alors de la source d'apparition du biais et de la possibilité d'accéder ou non aux variables sensibles concernées. A partir de là, il est possible de formaliser le problème : le point critique consiste à sélectionner les critères d'équité qui seront considérés pour la mitigation du biais. Plusieurs propriétés mathématiques d'un modèle prédictif ont été introduites, tel que demographic parity, equalized odds, equal opportunity, ou encore lack of disparate treatment, qui mesurent différemment la manière dont la distribution des prédictions doit rester la même entre les groupes sensibles ou non, et selon la réponse prédite. Ces propriétés ne sont pas forcément équivalentes et nécessitent un choix politique de traitement du biais, et donc un arbitrage entre performance et éthique (del Barrio, Gordaliza, & Loubes, 2020)

Les méthodes proposées pour corriger les biais peuvent se regrouper selon les 3 types d'intervention différentes :

En amont, par une opération de prétraitement des données, qui peut consister à enlever les variables sensibles, ce qui est le plus souvent insuffisant en raison de la possible dépendance entre la variable sensible et d'autres variables non-sensibles du modèle (qui peuvent agir comme un proxy). Il est également possible de retirer ou de pondérer les observations pour permettre de corriger un biais de représentation, ce qui revient à dire qu'il y a une différence entre la distribution des données d'entraînement et la distribution des données de test (ou d'usage). Les méthodes de repondération consistent à exploiter une connaissance du mécanisme de biais (biais de sélection, biais de censure) au travers d'une information auxiliaire et qui permet de construire les poids et de minimiser une fonction de coût pondérée par des techniques de type Horvitz-Thompson (utilisées couramment en sondage), (Ausset, Clémençon, & Portier, 2019).

Par la modification des algorithmes eux-mêmes en introduisant typiquement des contraintes d'équité sous forme de régularisation dans la fonction de coût. Dans le cas de la classification binaire par exemple, les notions de "disparate impact" se prêtent mathématiquement bien à une intégration dans la fonction (i.e peuvent se ramener à des problèmes convexes), pour laquel nous pouvons obtenir des modèles sans disparate impact (la probabilité de prédire la classe positive est indépendante des valeurs que peuvent prendre la variable sensible), (Zafar, Valera, Gomez-Rodriguez, & Gummadi, 2019). Des corrections de post-processing sont proposées pour faire correspondre la distribution des réponses dans les différents sous-groupes définis par la variable sensible. Ceci est fait notamment par des techniques de transport optimal (Chzhen, Denis, Hebiri, Oneto, & Pontil, 2020).

Robustesse

La robustesse d'un algorithme ou d'une décision est une propriété qui correspond à plusieurs problématiques différentes mais qui consiste à considérer l'impact sur les prédictions de l'algorithme des fluctuations des données d'apprentissage, et plus généralement, de la loi de probabilité des données d'apprentissage.

La robustesse est un sujet classique en statistique qui adresse la résistance des procédures statistiques à la contamination des données d'apprentissage par des données atypiques (outliers), c'est-à-dire lorsque la distribution d'intérêt est contaminée par une autre distribution. Ces approches sont assez classiquement traitées par l'utilisation de fonctions de coût robustes (M-estimateur), ce qui est la base et couramment utilisé en machine learning.

Cependant, ces approches permettent de garantir la robustesse uniquement à un certain types de perturbations. 'utilisation du machine learning sur des données et des phénomènes complexes nécessite une robustesse à des perturbations plus variées ou plus importantes, et noyées dans de grandes quantités de données. Pour cette raison, d'autres principes d'apprentissage sont introduits tels que la minimisation de distance entre modèles (des mesures de probabilités) à l'aide de plongements (embeddings) dans des espaces appropriés - Maximum Mean Discrepancy" (Alquier & Gerber, 2020) ou encore les "Median-of-Means" (Lecué & Lerasle, 2020) qui introduit un critère minimax. Enfin, le lien avec les modèles génératifs adversariaux (f-GAN) est aussi étudié (Gao, Liu, Yao, & Zhu, 2019).

Cette notion de robustesse que l'on peut rapprocher d'une certaine façon à une erreur de modèle (la contamination pouvant être importante), est complétée par une robustesse, presque plus classique aux variations de l'ensemble d'apprentissage ainsi qu'aux différents processus, parfois stochastiques, qui interviennent dans l'apprentissage. Il s'agit alors d'estimer l'incertitude d'une procédure de machine learning, ce qui est notoirement plus difficile que pour des modèles statistiques.

Cette notion de robustesse peut alors être la stabilité d'un algorithme dans le sens développé par (Yu & Kumbier, 2020) qui met en avant le cadre de "predictability, computability, stability" pour aboutir à une science des données véridique. L'introduction de perturbations multiples dans les données, le ré-échantillonnage, l'optimisation, etc. permet de réduire et d'estimer aussi la variance et l'erreur de prédiction. Cependant les approches classiques (validation croisée, ensemble de test) ne permettent d'avoir qu'une estimation globale de l'incertitude et de l'erreur de prédiction à attendre et peuvent dissimuler des situations très variées sur la qualité des prédictions individuelles. Ainsi, il paraît nécessaire de s'intéresser aussi à la variance des prédictions individuelles, ou à la construction d'un intervalle de prédiction. De telles mesures d'incertitudes pour des modèles prédictifs complexes existent pour les Random Forests (Infinitesimal Jackknife) mais aussi de manière générique avec l'utilisation des méthodes de prédiction conforme, ou qui en sont directement inspirées, telles que le Jackknife + (Barber, Candes, Ramdas, & Tibshirani, 2021). Ces dernières peuvent être utilisées pour construire des intervalles de prédiction en faisant très peu d'hypothèses sur les données et l'algorithme. Notablement, la garantie de la probabilité de couverture des intervalles est assurée grâce à la stabilité de l'algorithme.

Intelligibilité

L'intelligibilité de l'IA est une discipline qui connaît une forte croissance, parallèlement à celle de l'éthique. Elle prend d'une certaine façon le relais des questions classiques autour de la mesure de l'importance d'une variable dans un modèle de machine learning, tel que l'attribution d'un poids, et de l'analyse de sensibilité. Elle pose aussi de nouvelles questions comme celle de l'intelligibilité locale. Dans ce dernier cas, il ne s'agit pas de déterminer les variables importantes pour le modèle (et donc en moyenne pour l'ensemble des données d'apprentissage) mais d'identifier les variables importantes pour la prédiction faite pour un individu en particulier.

De nombreux concepts, mesures et graphiques ont été proposés pour répondre à ces questions, et un point important consiste à savoir si ces outils sont dépendants explicitement du modèle, ou s'ils sont "agnostiques du modèle". Cette dernière famille connaît le plus grand développement car elle permet d'aller vers une comparaison des explications : ces méthodes évitent d'expliquer la prédiction par le "comment" du calcul (les coefficients d'une régression, les nœuds d'un arbre), le plus souvent en se ramenant à des règles ou des modélisations plus simples. Un panorama assez complet est fourni dans (Molnar, 2021) en proposant une organisation selon des modèles interprétables par nature (modèle linéaire généralisé (GLM), modèle additifs (GAM), les arbres et règles de décision, les méthodes agnostiques, et enfin les méthodes d'explication fondées sur les exemples tels que les exemples contrefactuels et adversariaux.

Essentiellement, l'attribution d'importance globale exploite les méthodes classiques de "Permutation Feature Importance" pour évaluer la contribution d'une variable à la performance globale du modèle. L'utilisation de modèles approchés "interprétables" est aussi courante.

Dans le cas de l'intelligibilité locale, les méthodes agnostiques comme les méthodes graphiques (Partial Dependence Plot, Individual Conditional Expectation...) permettent de visualiser la sensibilité d'un modèle mais ces approches sont limitées à de petites dimensions. Pour considérer un plus grand nombre de variables, l'utilisation de modèles approchés locaux est alors recommandée : c'est la méthode LIME (Local interpretable model-agnostic explanations) (Ribeiro, Singh, & Guestrin, 2016). Cependant, la qualité ou la pertinence de l'approximation peut être discutable. Les valeurs de Shapley sont aussi très couramment utilisées pour attribuer une importance à toutes les variables en se fondant sur une décomposition additive de la prédiction. Cet indicateur est implémenté dans une librairie SHAP (Lundberg & Lee, 2017) et est présenté comme satisfaisant des propriétés mathématiques intéressantes pour l'intelligibilité, issues de la théorie des jeux coopératifs dont provient initialement le concept de valeur de Shapley. Avec l'utilisation importante de cet outil, des limites apparaissent aussi sur les approximations nécessaires aux calculs des valeurs de Shapley, la prise en charge des variables catégorielles ou leur interprétation en tant que valeur d'importance d'une variable (Amoukou, Brunel, & Salaün, 2021) Un axe de recherche est aussi le lien entre ces indicateurs et la causalité qui est également un sujet très actif (Schölkopf, 2019).

Contrairement aux approches précédentes, l'estimation de structures causales dans les modèles de machine learning suppose de modifier les modèles, ou les techniques d'estimation. En effet, l'identification de relations causales, et plus seulement corrélatives entre les variables, se fait notamment par l'emploi de modèles graphiques bayésiens ou structuraux. Si l'étude de la causalité est un objectif prioritaire, des critères dédiés à la performance prédictive n'ont pas forcément de sens. Comme par exemple, l'étude des « outcomes » potentiels en inférence causale, pour lesquels les objectifs d'intérêt sont l'effet causal moyen pour mesurer l'efficacité d'un traitement.

Nous pouvons aussi remarquer que la notion de causalité introduite par Pearl est parfois trop exigeante et celle-ci peut être relaxée en mettant en avant le lien avec la robustesse, la stabilité et la causalité (Bühlmann, 2020).
Enfin, nous évoquons aussi les approches bayésiennes qui permettent de traiter les problèmes de choix de modèle (avec les facteurs de Bayes) et de modélisation graphique et causale, et qui est une méthode de choix pour l'instant pour l'estimation de l'incertitude dans les réseaux de neurones via interprétation bayésienne du dropout (Wang & Yeung, 2020).

Cycle de vie du modèle

Le cycle de vie de modèle est à la croisée des enjeux d'ingénierie de l'IA et des problématiques de mise en production des modèles de data science et machine learning. Ces nouvelles problématiques induisent l'émergence de nouveaux métiers tels que le Machine Learning Engineer, au côté du Data Scientist, qui est en charge de l'exploitation du modèle. Il s'agit de faire face à la dérive temporelle des modèles. Il existe deux types de risques liés à l'obsolescence des données ayant servi à entraîner le modèle : les dérives virtuelles et les dérives réelles (Gama, Zliobaite, Bifet, Pechenizkiy, & Bouchachia, 2014)

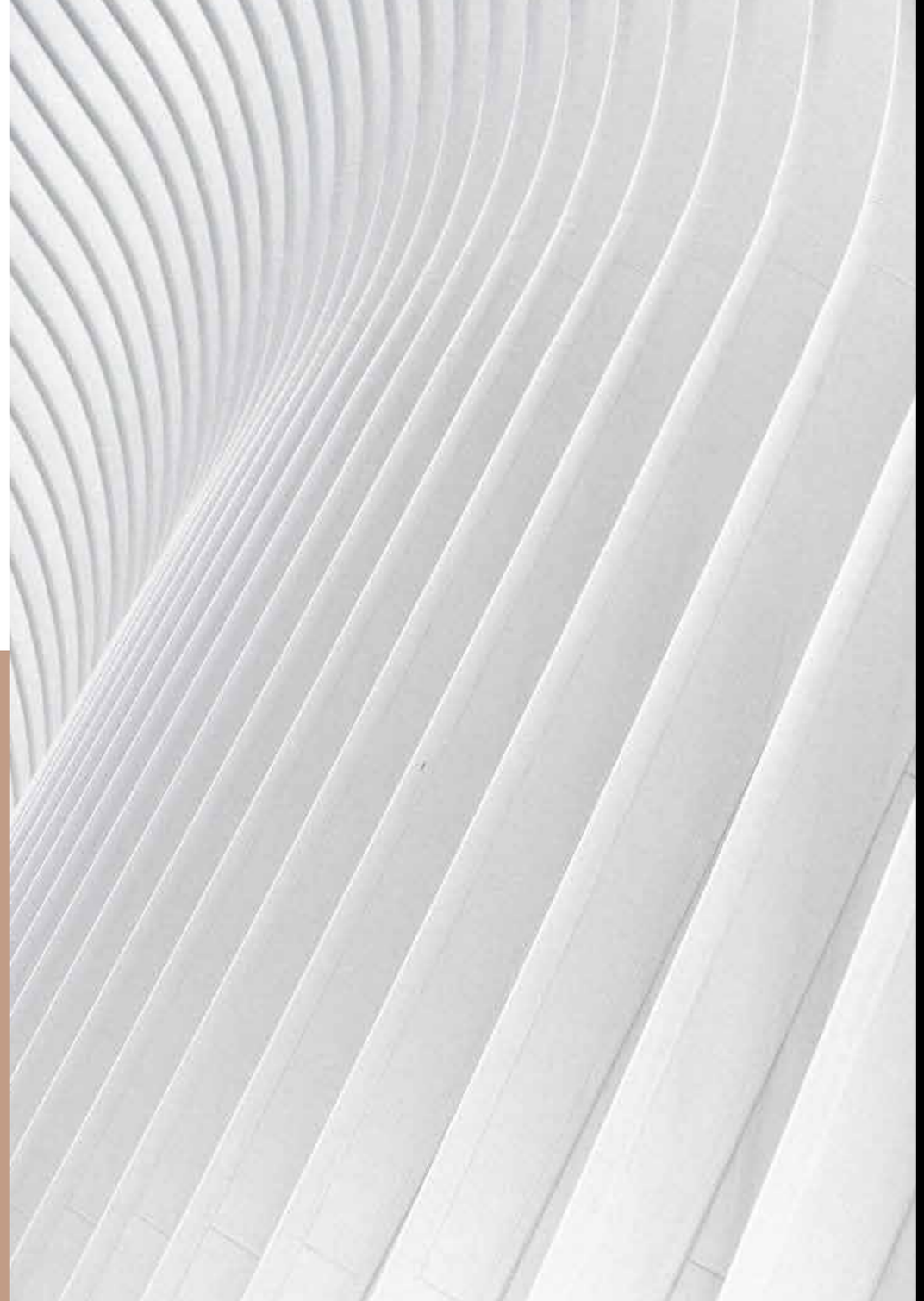
Une dérive virtuelle survient lorsque la distribution des données en entrée (variables explicatives) change, sans que la distribution de la cible connaissant ces variables n'ait varié. Plus généralement, on peut avoir un "covariate shift", c'est-à-dire, une modification de la distribution des variables d'input tout en gardant la même relation entre la variable de sortie et les inputs. Au contraire, une dérive réelle se définit par un changement de lien causal entre variables explicatives et cible. Dans ce cas, il est nécessaire de retirer les données les plus obsolètes qui ne sont plus représentatives du contexte de production, et bien sûr si possible, de les remplacer par des données plus récentes. On risque ici une chute de performance : c'est par exemple le cas en détection de fraude lorsqu'un fraudeur change de stratégie (Dal Pozzolo, Boracchi, Caelen, Alippi, & Bontempi, 2018). Afin de prendre les décisions les plus pertinentes, on s'appuie sur deux types de monitoring : supervisé et non supervisé. Le monitoring supervisé consiste à mesurer la performance réelle du modèle et détecter les dérives réelles (Truong, Oudre, & Vayatis, 2020), mais il nécessite de disposer des labels réels et peut manquer de réactivité (voire être impossible, pour un octroi de crédit par exemple). On peut parfois pallier ce problème en accélérant la labellisation de tout ou partie des données (Aggarwal, Kong, Gu, Han, & Yu, 2014). Le monitoring non supervisé surveille les variations de la distribution des données d'entrée (Székely & Rizzo, 2017), et peut donc se faire en temps réel. En contrepartie, il faut déterminer si une dérive observée de ces données est liée à une dérive réelle, et si oui, sur quel historique le modèle doit être ré-entraîné.

Le stacking de modèle et le suivi de modèles entraînés sur des horizons différents offrent une flexibilité et une réactivité grâce à la mise en production de celui qui a les meilleures performances dans un contexte donné (Gomes, et al., 2017). Une alternative est l'adaptation de domaine qui consiste à trouver une transformation qui ramène les nouvelles données sur l'ensemble d'entraînement d'une manière qui n'affecte pas les labels (Kouw & Loog, 2021).

Enfin, les pertes de performance liées à la dérive des données peuvent être anticipées à l'aide de l'out-of-time testing qui consiste à réaliser une validation croisée en contraignant l'ensemble d'apprentissage à être situé dans le passé de l'ensemble de test. On peut ainsi détecter dès le développement du modèle que les données dérivent dans le temps et mettre en place des solutions de monitoring et d'alerting appropriées.

Cependant, le réentraînement peut augmenter les risques dus à des problèmes de qualité de nouvelles données. Une première technique permettant de dé-risquer la mise à jour d'un modèle, est l'A/B testing (Kohavi, Deng, & Walker, 2013). Dans le cas du canary testing par exemple, on applique les recommandations du nouveau modèle pour, par exemple 1% des données entrantes et on garde l'ancien pour le reste des données. On peut alors comparer les résultats des deux versions dans les mêmes conditions avant de basculer définitivement.

Une alternative est le shadow deployment qui consiste à déployer les deux modèles simultanément tout en suivant les recommandations de l'ancien modèle. On prend donc moins de risque mais on dispose d'une évaluation moins fiable du nouveau modèle.



CONCLUSION

CONCLUSION

Nous l'expérimentons chaque jour, l'IA est aujourd'hui omniprésente dans nos vies professionnelles comme personnelles et dans tous les domaines d'application dont certains domaines « à risque » comme la santé ou les services publics. Comme pour toute technologie nouvelle, l'utilisation de l'IA entraîne des risques aisément identifiables mais aussi, de par sa nature abstraite, des risques sous-jacents difficilement détectables ou mesurables. Ce sont ces risques dont l'UE souhaite protéger ses citoyens par l'intermédiaire d'un texte définissant un cadre réglementaire, et d'une définition commune et explicite de l'IA de confiance. Cette démarche réglementaire possède une vertu immédiate qui est celle de la sensibilisation de l'opinion publique à l'enjeu que représente une adoption éclairée de l'IA, afin de donner confiance à la société et aux entreprises dans son utilisation.

Ainsi, nous sommes convaincus chez Quantmetry que cette réglementation sera un vecteur d'une IA mieux appréciée, valorisée et maîtrisée.

La mise sous contrôle des applications dites « à risque » aura par ailleurs un double effet positif : évidemment en sécurisant les IA ciblées directement mais également, par effet d'entraînement, en élevant le niveau d'exigence et de transparence pour l'ensemble des algorithmes mis en œuvre.

L'application concrète du cadre général défini par l'UE reste toutefois un réel challenge pour les entreprises car la déclinaison des concepts et des directives est aujourd'hui soumise à interprétation : c'est à présent aux institutions sectorielles et aux acteurs spécialisés de préciser les modalités d'application. En effet, des exigences trop complexes ou trop administratives qui viendraient s'ajouter à des contraintes d'application déjà existantes (notamment en milieu régulé) rendraient alors la mise en œuvre de l'IA non viable et seraient ainsi un frein à son adoption. Un autre risque identifié est la possible surenchère de communication ou de labels à ce sujet qui aboutirait à un phénomène « d'Ethic-Washing » en décrédibilisant une démarche bénéfique.

En effet, même s'il n'existe pas aujourd'hui de consensus sur les méthodes et les techniques à mettre en œuvre pour implémenter une IA de confiance, des outils existent comme par exemple des questionnaires pour auditer les algorithmes ou des méthodes basées sur les valeurs de Shapley pour améliorer l'intelligibilité des modèles d'IA. Ces outils sont par nature imparfaits, l'état de l'art scientifique et méthodologique continuant à évoluer, mais apportent dès aujourd'hui un premier niveau de réponse à cette problématique.

L'ensemble des acteurs de l'IA, du concepteur à l'utilisateur, gagneront à sensibiliser l'opinion sur l'importance du sujet de l'IA de confiance, à être attentifs aux évolutions des techniques applicables et les entreprises trouveront un avantage certain à s'engager dès à présent dans la mise en œuvre d'une IA de confiance.

BIBLIOGRAPHIE

BIBLIOGRAPHIE

Aggarwal, C. C., Kong, X., Gu, Q., Han, J., & Yu, P. S. (2014). Active learning: A survey. Data Classification: Algorithms and Applications , 571-605.

Alquier, P., & Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy.

Amoukou, S. I., Brunel, N. J.-B., & Salaün, T. (2021). The Shapley Value of coalition of variables provides better explanations.

Ausset, G., Clémençon, S., & Portier, F. (2019). Empirical Risk Minimization under Random Censorship: Theory and Practice.

Bühlmann, P. (2020). Invariance, Causality and Robustness (with discussion). Statistical Science 35, 404-426.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. Annals of Statistics, 49(1), 486-507.

Carpentier, L. (2021). L'algorithme, nouvelle machine à tubes. Vraiment ? Le monde.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair Regression with Wasserstein Barycenters.

Croak, M. (2021). Récupéré sur blog.google: https://blog.google/technology/ai/marian-croak-responsible-ai/

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems, 29-8, 3784-3797.

Dastin, J. (2018). Récupéré sur reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

del Barrio, E., Gordaliza, P., & Loubes, J.-M. (2020). Review of Mathematical frameworks for Fairness in Machine Learning.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. ACM Comput. Surv. 46, 4, Article 44 .

Gao, C., Liu, J., Yao, Y., & Zhu, W. (2019). Robust Estimation Via Generative Adversial Networks. International Conference on Learning Representations. New Orleans.

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., . . . Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. Mach Learn 106, 1469–1495.

Kohavi, R., Deng, A., & Walker, T. (2013). mproving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data. Sixth ACM International Conference on Web Search and Data Mining, (pp. 123–132). New York, NY, USA.

BIBLIOGRAPHIE

Kouw, W. M., & Loog, M. (2021). A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 766-785.

Lecué, G., & Lerasle, M. (2020). Robust machine learning by median-of-means : theory and practice. Annals of Statistics, 48(2), 906-931.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS).

Molnar, C. (2021, 12 4). Interpretable Machine Learning A Guide for Making Black Box Models Explainable. Récupéré sur Github: https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Schölkopf, B. (2019). Causality for Machine Learning.

Székely, G. J., & Rizzo, M. (2017). The Energy of Data. Annual Review of Statistics and Its Application, 4-1, 447-479.

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. Signal Processing, 167.

Vakulina, O. (2019). Euronews. Récupéré sur https://fr.euronews.com/2019/04/03/des-milliards-d-euros-pour-l-intelligence-artificielle-europeenne

Wang, H., & Yeung, D.-Y. (2020). A Survey on Bayesian Deep Learning. ACM Computing Surveys (CSUR), 53(5), 1-37.

Yu, B., & Kumbier, K. (2020). Veridical data science. Proceedings of the National Academy of Sciences, 117-8, 3920-3929.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. Moritz Hardt.

CONTACT

Guillaume BODIOU
Partner
gbodiou@quantmetry.com

Amélie SEGARD
Data Scientist Senior - Reliable AI
asegard@quantmetry.com

Martin LE LOC
Directeur - Reliable AI
mleloc@quantmetry.com

Tsvetina BACHEVA
Manager - AI Strategy
tbacheva@quantmetry.com

Jonathan CASSAIGNE
Directeur - AI Strategy
jcassaigne@quantmetry.com