N°4 | STEPWISE

Quantmetry

RÉDACTEURS

Rémi Adon, Marc Artaud de la Ferrière, Maya Azouri,
Adèle Guillet, Guillaume Hochard, Antoine Isnardy,
Stéphane Jankowski, Grégoire Martinon, Aurélia Nègre,
Amélie Segard, Pablo Valverde, Julien Vong

SOMMAIRE

PRÉFACE

L'intelligence artificielle (IA) est généralement définie comme un ensemble de concepts et de technologies mises en œuvre en vue de réaliser des machines capables de reproduire le comportement humain. Au-delà du champ scientifique, l'IA constitue à présent un paradigme philosophique. Qu'est-ce que l'IA, sinon une tentative pour libérer l'être humain de travaux difficiles et répétitifs, et lui permettre de se concentrer sur des tâches à forte valeur ajoutée - mais aussi des tâches qui lui apportent une plus grande satisfaction personnelle ?

Surtout, l'IA est devenue un secteur économique extrêmement dynamique. Technologies en place ou à prévoir, cas d'usage envisageables, données à disposition, aspects réglementaires... Autant d'éléments à prendre en compte lorsqu'on s'attaque à ce domaine passionnant, où une réactivité immédiate est souvent d'importance capitale.

La démarche agile et toutes les méthodes de gestion de projet qui y sont associées apparaissent de plus en plus indispensables ; elles ont le fort mérite de mitiger le risque d'aboutir à un résultat décevant, voire inutilisable. Cependant, cette agilité peut compliquer les grands choix stratégiques, qui nécessitent de la prise de hauteur. Cette réflexion se nourrit, entre autres, de questionnements épistémologiques majeurs sur la robustesse des décisions sous-traitées à des outils d'IA ; jusqu'où l'automatisation est-elle ou sera-t-elle possible, et jusqu'où peut-on accepter qu'elle le soit ?

L'un de ces questionnements porte donc, naturellement, sur la confiance que l'on peut accorder à ces outils, et de façon symétrique, sur celle que l'on peut accorder aux *actants* - ces hommes et femmes qui ont à entériner les ensembles croissants de décision proposés par les IA. Cette question joue un rôle central dans l'avancée des sciences et techniques, depuis l'invention du

premier dispositif auquel la gestion des conséquences d'une décision risquée a été déléguée. L'intelligibilité d'une décision constitue sans doute un facteur puissant de confiance, même si dans notre vie quotidienne, nous faisons déjà spontanément confiance à des mécanismes et algorithmes qui nous accompagnent depuis longtemps (le fonctionnement d'une voiture, la capacité d'un distributeur à nous fournir de l'argent, l'ouverture des portes d'un métro automatique…), sans chercher à comprendre la façon dont ils fonctionnent.

Toutefois, le nouveau paradigme philosophique sous-jacent à l'émergence d'algorithmes capables d'accélérer notablement des pans entiers de l'économie, de l'industrie ou de la vie personnelle, en proposant de remplacer des systèmes où l'humain reste encore le décideur principal, repose plus âprement cette question de la confiance et de l'intelligibilité des outils d'IA : peut-on confier qui notre économie, qui notre vie à une machine virtuelle ? Pouvons-nous accepter de ne pas tout comprendre de ses rouages ? Cela d'autant plus à une époque où l'information - mais pas la connaissance - est d'accès quasi-illimité ? Dans un contexte où de nombreux outils d'IA sont élaborés et calibrés en tirant parti de corrélations mises en lumière entre des données massives, l'intelligibilité que nous recherchons s'adresse aussi à ces choix de données ; est-on bien sûr que celles-ci sont *représentatives* ?

Notre position de pure player sur les sujets data nous oblige à appréhender ces interrogations et adopter une démarche visionnaire ; par le biais d'une sémantique clarifiée, il nous faut apporter des réponses pratiques à nos clients. Il s'agit pour nous de bâtir et perpétuer une relation de confiance, tout autant que d'accroître notre excellence technique et stratégique sur l'utilisation des données. C'est pour cette raison qu'une partie conséquente de notre activité porte sur des sujets de stratégie data. Notre activité de R&D est ainsi fortement consacrée à ce type de réflexion. Elle nous permet - nous impose même - de considérer des aspects importants de nos missions - telle l'intelligibilité des outils et des données - dont les dimensions dépassent celles de la simple réalisation de missions techniques.
Cette réflexion nous est utile en mission, afin d'alerter nos clients sur les risques cachés et les pièges. Mais comme nous considérons que ces aperçus puissent avoir un intérêt plus général, nous dédions aussi une partie de notre activité à présenter ces sujets en public lors de conférences et de meetups,

ainsi qu'à rédiger des documents destinés à différentes publics (articles de blogs, dictionnaires, livres blancs ...)

En parcourant nos publications, les sujets qui nous tiennent à cœur émergent clairement : comment arriver à embaucher les bons collaborateurs dans un marché tendu et caractérisé par une pénurie de profils techniques (« Dessine-moi un data scientist », « Dessine-moi un data engineer ») ; quelles précautions adopter pour éviter qu'un projet data en reste au stade de *proof of concept* et pour en faciliter la mise en production (« Y a-t-il une vie après le POC ») ; comment valider certains éléments de langage communs, indispensables pour faciliter les échanges entre startups et grands groupes (« Enigma Vol 1. Quand startup et grand groupe ne parlent pas la même langue »).

La nécessité de pouvoir bien interpréter les modèles d'IA, leur élaboration, leur périmètre de pertinence et les résultats de leur mise en œuvre font donc partie de ces sujets « de cœur » et d'importance[1] , pour de multiples raisons qui seront explicitées dans ce livre blanc. Nous avons la conviction que l'importance de ce sujet ne fera que croître dans les prochaines années, et nous espérons que les idées que nous développons ici trouverons à résonner dans vos activités professionnelles actuelles et futures. Accompagnez-nous, au travers des pages qui suivent, à la découverte d'une problématique majeure pour notre société de plus en plus technologique.

Bonne lecture !

Nicolas Bousquet, PhD
Directeur scientifique

Alberto Guggiola, PhD
Senior Data Scientist

[1] Sur le même sujet, voir aussi l'article - accepté par la Revue d'Intelligence Artificielle - Pégny, M. and Ibnouhsein, M.I., 2018. Quelle transparence pour les algorithmes d'apprentissage machine?

REMERCIEMENTS

La rédaction de ce livre blanc a été menée en s'appuyant sur les retours d'expérience des entreprises en pointe sur les sujets Big Data. Aussi, nous alternons parties explicatives et extraits d'interviews qui viennent illustrer le propos.

Nous tenons à remercier chaleureusement GRTgaz, BPCE, les Hôpitaux Universitaires de Strasbourg et PSA pour avoir consacré du temps à ce livre blanc.

Le département D2AL de GRTgaz
La Direction Achats Approvisionnements Logistique de GRTgaz (D2AL) créée en 2011 a pour objectif de répondre à trois enjeux majeurs sur les achats qui concourent directement à la maîtrise industrielle, au développement du réseau et aux enjeux de sûreté de GRTgaz ainsi qu'à son ambition de devenir le transporteur de gaz de référence en Europe. Accompagnée par le Datalab depuis 2016, la D2AL se dote notamment d'outils de machine learning pour optimiser les stocks des entrepôts.

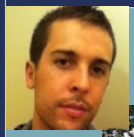
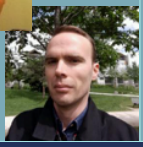
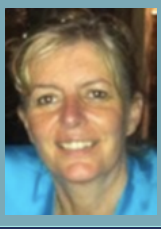
La direction des risques, de la conformité et des contrôles permanents de BPCE
Au sein de ce département, les équipes travaillent sur les risques, la conformité, les contrôles permanents et la sécurité pour l'ensemble du groupe BPCE (organe central commun à la Banque populaire et à la Caisse d'épargne française).

Les Hôpitaux Universitaires de Strasbourg
Le CHU de Strasbourg est un hôpital de secteur appelé à assurer les soins courants à la population de Strasbourg et de ses environs. Le CHU de Strasbourg est aussi hôpital d'appel : compte tenu de son équipement de pointe et de son caractère universitaire, il est destiné à recevoir également les malades de secteurs géographiques éloignés que les centres hospitaliers généraux, non

dotés des mêmes équipements, ne peuvent prendre en charge. Deux projets de recherche en sénologie ont été lancés, en collaboration avec Quantmetry afin de mieux pouvoir cerner les possibilités d'apparition du cancer du sein et ses complications.

PSA

PSA a implanté sa Data Factory à Poissy en 2018, une entité à part entière permettant de faire de l'exploitation des données un nouveau levier de performance. La Data Factory est composée d'experts de la donnée qui accompagnent les métiers dans la recherche d'amélioration de leur performance opérationnelle comme par exemple l'optimisation des ateliers de peinture.

GRTGAZ, D2AL	BPCE, RISQUES, CONFORMITÉ ET CONTRÔLES PERMANENTS	PSA, DATA FACTORY	CHU DE STRASBOURG
OLIVIER LE BIHAN, *Technicien Logistique Senior* ET **VINCENT FOUCHÉ,** *Approvisionneur*	**BIBI NDIAYE,** *Responsable des Modèles Internes & Big Data*	**EMMANUEL MANCEAU,** *Responsable Projets et Stratégie Data*	**CAROLE MATHELIN,** *Responsable de l'unité de sénologie,*
 			

MIEUX COMPRENDRE L'IA POUR INSTAURER UN CLIMAT DE CONFIANCE

POURQUOI UN LIVRE BLANC SUR L'INTERPRÉTABILITÉ DES MODÈLES ?

L'intelligence artificielle (IA) prend de plus en plus sa place au centre du débat public. Des articles de presse ou des reportages au JT nous présentent quotidiennement des sujets plus ou moins liés à l'IA. Étrangement, ils alternent de manière un peu schizophrène deux approches opposées. D'un côté, une confiance absolue dans un avenir où toutes nos tâches seraient facilitées par l'IA, qui permettrait aussi de créer de nouvelles opportunités à n'en plus finir ; d'un autre, une peur généralisée (et un peu irrationnelle, même si parfois justifiée par des événements de l'actualité) qui semble avoir comme seul point de chute possible une fuite de toute avancée dans ce domaine.

Chez Quantmetry, nous avons la chance de traiter ces sujets chaque jour, et donner notre point de vue sur ces thèmes via un livre blanc nous semble une bonne manière de déclencher une discussion plus sereine au sein de la communauté (et pas uniquement entre *data scientists*[2]), de donner le juste poids aux craintes et aux espoirs, de partager l'impact que l'intelligence artificielle a déjà et pourra avoir dans le futur sur le monde du travail, sur la santé économique des entreprises françaises et non sur la société dans son ensemble.

Plusieurs axes d'analyses seraient intéressants à développer : quelles technologies ou modèles ont fait leurs preuves récemment, et dans quels contextes ? Quels cas d'usage ont émergé dans les différents secteurs métier, et quelles similarités ont pu être identifiées entre eux ? Quelles contraintes faut-il considérer au moment de structurer une organisation pour permettre à l'IA d'être déployée le plus efficacement possible au sein d'une entreprise ? Quelles contraintes légales prévoir pour l'utilisation de ces algorithmes ?
Bien entendu, loin d'être des îlots indépendants, ces thèmes sont destinés à se croiser. Nous nous sommes aperçus au fil de l'eau que l'interprétabilité des

algorithmes et des modèles, objet principal de ce livre, nous oblige à réfléchir à toutes les questions ci-dessus, et à d'autres encore. Mais pourquoi cette notion d'interprétabilité est-elle donc si cruciale ?

QUARANTE-DEUX

Dans le célèbre œuvre de science-fiction de Douglas Adams « Le guide du voyageur galactique [3] », des chercheurs construisent Deep Thought, un ordinateur super puissant, afin de lui poser la « grande question sur la vie, l'univers et le reste ». Celui-ci, après un calcul d'à peine sept millions et demi d'années, leur donne une réponse toute simple et dont il assure l'exactitude : quarante-deux.

Cet épisode est très connu dans la communauté nerd du monde entier (et donc, par association, à de nombreuses personnes qui travaillent dans le monde de la data) : il pointe du doigt le risque d'une hyper confiance dans les capacités des ordinateurs à trouver les réponses aux questions qui nous troublent. L'auteur pense que, même si dans un futur lointain ce souhait pourra être réalisé, les ordinateurs trouveront des réponses « compréhensibles » par eux, mais qui auront peu ou pas de sens pour nous.

Est-ce que cette crainte est justifiée ? Les choses ont-elles bougé depuis les années 70, date de sortie du livre ? Y a-t-il vraiment le risque d'être en route vers une intelligence artificielle qui nous donnera des réponses incompréhensibles et donc difficilement utilisables ?

UN CHANTIER NÉGLIGÉ À L'AVANTAGE DE LA RECHERCHE DE LA PERFORMANCE

Même si la vision d'Adams reste exagérée, ce risque existe bel et bien. Cela dépend en partie du parcours historique du Machine Learning, et de la nature des premières tâches qu'on chercha à automatiser grâce à lui. Au final, à qui importe le processus de décision utilisé par une machine afin de correctement

identifier un chiffre 7 écrit à la main et de le distinguer d'un 4 ? Du moment que ça marche…

[3] D. Adams, Le Guide du *voyageur galactique,* Gallimard 2005

En suivant ce principe, la communauté scientifique visa pendant des décennies l'amélioration des performances (temps nécessaire pour donner une réponse, fiabilité des prédictions, convergence des algorithmes) plutôt que la « lisibilité » de la démarche. Mais avec le déploiement de méthodes de Machine Learning dans des nouveaux contextes métier, cette hypothèse de travail changea : un patient (ou même un médecin) ne considérerait pas comme un petit cadeau sacrifiable la compréhension du rationnel derrière la préconisation d'une chimiothérapie par une IA.

Qui plus est, une réponse transparente pour un mathématicien ne l'est pas forcément pour un opérationnel ; cela devient encore plus remarquable dans l'âge des Big Data. Prenons le cas extrême d'un modèle déterministe avec des centaines de variables d'entrée : pour un mathématicien, son interprétabilité est assurée (pour chaque input nous pouvons calculer exactement la sortie qui sera produite par l'algorithme) ; un expert métier, au contraire, aura l'impression de se noyer parmi toutes ces variables, et au final ne pourra qu'utiliser le résultat sans le comprendre, ou bien ne pas le considérer du tout.

Cet exemple démontre aussi que même l'hypothèse de l'existence d'une tension entre explicabilité et performance, qui impliquerait la nécessité de chercher un compromis entre les deux, n'est pas vraiment justifiée : il existe des algorithmes simples (et peu performants) qui sont difficilement interprétables, aussi bien qu'il existe des outils et des *frameworks* qui nous permettent de bien comprendre le fonctionnement et les sorties des méthodes les plus sophistiquées.

UNE PRIORITÉ RECONNUE

Le rapport « Donner du sens à l'intelligence artificielle [4] », rédigé par le mathématicien et député Cédric Villani et rendu public en mars 2018, vise entre autres à « proposer quelques pistes permettant de poser les bases d'un cadre éthique pour le développement de l'IA et à faire vivre ce débat dans la société.

[4] Villani, C., Schoenauer, M., Bonnet, Y., Berthet, C., Cornut, A.C., Levin, F., Rondepierre, B. and Biabiaby-Rosier, S., 2018. Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne. *Rapport public, Premier ministre.*

Ce n'est pas par hasard que le premier parmi les piliers proposés concerne notre capacité à interagir avec les intelligences artificielles :

« En premier lieu, il faut accroitre la transparence et l'auditabilité des systèmes autonomes d'une part, en développant les capacités nécessaires pour observer, comprendre et auditer leur fonctionnement et, d'autre part, en investissant massivement dans la recherche sur l'explicabilité. »

Le manque d'interprétabilité est donc considéré comme l'un des facteurs bloquants l'industrialisation de l'IA. Mais pourquoi ?

Tout d'abord, de plus en plus de contraintes légales (avec l'arrivée du RGPD en Europe par exemple) donnent une réponse très directe à cette question : dans le futur, de nombreuses applications d'IA seront simplement bloquées en l'absence d'un niveau d'interprétabilité satisfaisant. Cela ne concernera pas que le domaine médical (dans lequel la vigilance est particulièrement stricte), mais tout secteur dans lequel des données personnelles sont susceptibles d'être traitées.

De manière plus générale, un manque d'interprétabilité génère des problèmes dans les différentes phases d'un projet data : un modèle black box rend difficile l'intégration d'intuitions métier, la compréhension de la dérive de ses performances dans le temps, la possibilité de modifier les actions prévues suite à ses sorties. L'interprétabilité nous aide à éviter des pièges qui pourraient être extrêmement dangereux en production.

UN DICTIONNAIRE PAS ÉVIDENT À CONSTRUIRE

Les quelques lignes tirées du rapport Villani contiennent un ensemble de mots-clés et de concepts que nous retrouverons tout au long du livre. Ce dictionnaire de termes n'est pas évident à construire, car des toutes petites nuances entrent en jeu, et une partie de subjectivité reste toujours dans chaque définition.

L'intelligibilité (ou de manière équivalente l'anglicisme « interpretabilité »), sujet principal de ce livre, désigne la « capacité à expliquer ou à présenter en des termes compréhensibles par un humain [5] ». L'intelligibilité d'un algorithme ou

[5] Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

d'un modèle désigne par extension le degré avec lequel on peut comprendre leur fonctionnement. Notons que cette définition est très loin d'une mesure objective et explicite, elle dépend fortement de la manière dont les humains raisonnent, comprennent et traitent l'information.
La notion d'explicabilité insiste davantage sur la capacité à expliquer et à être compris par un ensemble d'utilisateurs sans exiger des prérequis concernant le bagage éducatif. Cependant, les concepts d'intelligibilité et d'explicabilité restent proches et ne se distinguent que selon qu'ils insistent plus sur la capacité à comprendre que la capacité à expliquer. Notons que la maxime bien connue de Boileau vient rendre un peu plus floue la nuance entre les deux : « *Ce que l'on conçoit bien s'énonce clairement, Et les mots pour le dire arrivent aisément.* »
Le concept de transparence est plus large et a plusieurs significations. Il est tantôt employé pour insister sur des propriétés normatives des traitements de données (des propriétés dont il serait souhaitable qu'elles soient vérifiées : équité, loyauté, etc.) et tantôt pour décrire leurs propriétés (l'intelligibilité et l'explicabilité).
Le concept d'équité (*fairness* en anglais) désigne la propriété pour un traitement de ne pas induire d'effet discriminant à l'égard d'une catégorie particulière de la population.

Nous remarquons en passant qu'une vraie discussion autour de ce sujet ne peut pas se passer d'une réflexion plus large autour des sciences cognitives. Qu'est-ce que des mots comme « comprendre » ou « expliquer » peuvent vouloir dire ? Ce sont qui « les autres » qui doivent pouvoir comprendre nos explications ? Quelle est notre capacité à « interpréter » le monde autour de nous ? Est-ce qu'elle peut être définie une fois pour toutes, sans prendre en compte des aspects culturels, historiques, géographiques et sociaux ?

UN SUJET À ANALYSER SOUS PLUSIEURS POINTS DE VUE

Le sujet de l'interprétabilité des modèles est donc très riche, et peut être approché de plusieurs points de vue : c'est pour cela que nous avons décidé

d'axer nos réflexions autour de quatre principaux sujets qui forment les quatre chapitres de ce livre.

Nous commençons avec un focus sur l'impact que la demande d'une intelligibilité accrue aura dans le monde de l'entreprise, et sur les conditions à respecter afin de pérenniser les avantages qui seront générés par la révolution IA.

La nécessité de comprendre ce qui se passe derrière les algorithmes de *machine learning* ne concerne pas seulement les profils techniques : nous dédions ainsi un chapitre aux enjeux de ce sujet au sein de la société au sens large, avec une multitude d'acteurs concernés aux besoins variés et parfois contradictoires.

Ensuite, nous nous intéressons aux méthodes utilisables pour améliorer la compréhension des modèles d'IA en fonction de la typologie de données en entrée, ainsi qu'aux investissements à imaginer pour les mettre en production.

Enfin, nous vous partageons notre « vie de data scientist » au travers de quatre missions dans les secteurs bancaire, industriel, santé et ressources humaines, pour démontrer comment les questions d'intelligibilité se retrouvent de manière transverse dans les différents secteurs professionnels et chez tous les acteurs métiers.

ALBERTO GUGGIOLA

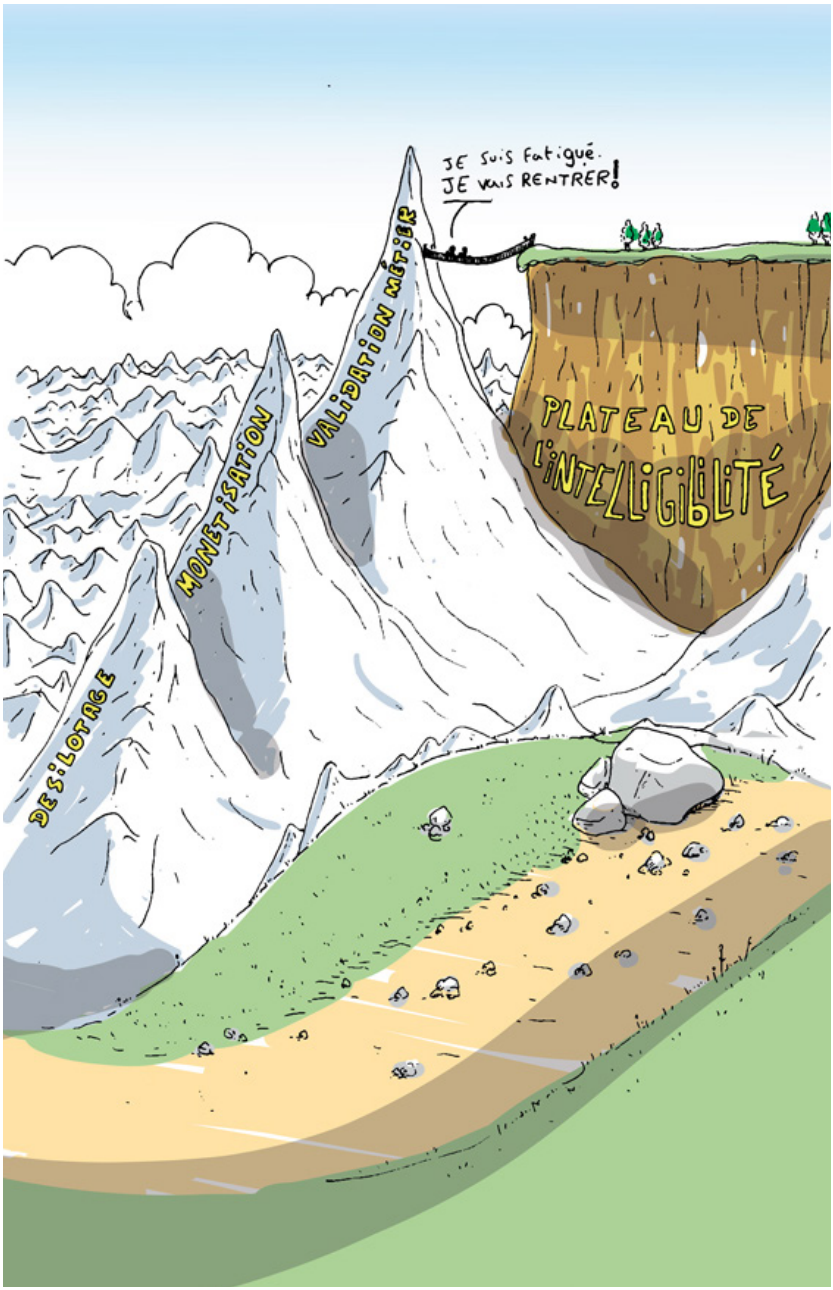
1. ENJEUX ET OPPORTUNITÉS POUR L'ENTREPRISE

Les révolutions technologiques obéissent souvent à un même modèle de diffusion de l'innovation similaire : touchant d'abord les spécialistes d'un secteur, puis se diffusant au grand public selon la célèbre courbe décrivant les principales populations : innovateurs, premiers adeptes, majorité précoce, majorité tardive et retardataires.

De par leur inertie, les entreprises (excepté les créateurs de l'innovation) appartiennent souvent à la catégorie des retardataires et sont donc les dernières à bénéficier pleinement des avantages offerts par la nouveauté. Ces révolutions nécessitent de changer les manières de travailler (organisation, outils...) et seuls des déclencheurs forts permettent de lancer la dynamique du changement : c'est ce que nous analyserons dans la première partie de ce chapitre, en présentant une **interprétabilité accrue comme le déclencheur de l'adoption de l'innovation data dans les entreprises.**

On peut cependant arguer que l'utilisation de la donnée, du chiffre, n'est pas nouvelle pour tous les secteurs puisque de nombreux modèles d'aide à la décision ou d'analyse statistique existent et sont déjà utilisés depuis longtemps ; nous traiterons donc dans une seconde partie **la spécificité des modèles de *data science* par rapport aux autres modèles statistiques qui préexistaient.** Nous détaillerons les raisons pour lesquelles la notion d'interprétabilité est nouvelle et propre à la révolution data.

Enfin, l'entreprise étant par essence une aventure humaine, nous détaillerons **les convictions qui nous animent en matière de gouvernance et d'organisation** : le déclencheur est une allumette, mais il faut savoir créer les bonnes conditions pour maintenir la flamme.


JE suis fatigué.
JE vais RENTRER!
VALIDATION MÉTIER
MONETISATION
DESILOTAGE
PLATEAU DE L'INTELLIGIBLITÉ

A. LE DÉCLENCHEUR DE L'ADOPTION DE L'IA EN ENTREPRISE

L'intelligence artificielle a aujourd'hui des impacts très visibles sur notre vie quotidienne, grâce à quelques grands acteurs (multiplication des assistants vocaux, habitude de consommation via les recommandations intelligentes d'achats…). **Dans le monde de l'entreprise en revanche, elle représente encore la matière première d'une transformation très profonde, dont le potentiel ne sera réellement activé que grâce à une interprétabilité accrue des modèles.**

IMPACT DES RÉVOLUTIONS PRÉCÉDENTES SUR NOS MANIÈRES DE TRAVAILLER

Considérons un parallèle avec deux situations du passé qui ont également eu un impact fort sur nos manières de travailler : la révolution industrielle d'une part et la récente digitalisation du traitement de l'information d'autre part. Dans ces deux révolutions, une technologie nouvelle se diffuse progressivement d'un secteur nouveau pour finalement toucher l'ensemble des acteurs de la sphère économique.

Lors de la première révolution industrielle, le perfectionnement de la machine à vapeur par James Watt fut très clairement une avancée majeure d'un point de vue technique et scientifique et impacta rapidement de nombreux secteurs : transport, manufacture, industrie. Mais ce n'est qu'**avec Taylor que ces évolutions techniques inspirèrent l'organisation scientifique du travail, qui changea le monde de l'entreprise dans son ensemble.**

Pour la révolution digitale, de la même manière, la plupart des applications utilisées encore aujourd'hui en informatique existent depuis plusieurs décennies, mais seule l'introduction d'interfaces d'utilisation plus *user friendly* banalisa au fur et à mesure ces compétences : désormais, une majorité de la population (peu importe le secteur d'activité) doit toucher à ces outils de manière quasi-continue. Ce changement de perspective produit également un effet sur le système éducatif, et les enfants commencent désormais à développer ces compétences (qui leur seront utiles dans tous les cas) dès l'école primaire.

Ces deux précédentes révolutions qui étaient originellement « technologiques » ont eu un impact très fort sur les manières de travailler et sur la productivité qui en découle grâce à des déclencheurs différents : une nouvelle organisa-

tion du travail pour la révolution industrielle et les interfaces utilisateurs pour la révolution digitale. Ces deux déclencheurs, bien que positifs pour la productivité n'ont pas échappé la critique : aliénation et perte de sens des ouvriers pour le Taylorisme ou pression accentuée sur les employés, difficulté de déconnecter et discerner les niveaux d'urgence pour la digitalisation.

OÙ EN EST L'IA DANS SA RÉVOLUTION ?

Comme évoqué en introduction, l'IA a déjà une longue histoire scientifique (recherches dans les bases mathématiques depuis les années 50, relance du sujet avec des ordinateurs de plus en plus puissants à partir des années 80) et des applications concrètes dans de nombreux secteurs (industrie, marketing, santé, retailing) depuis les années 2000 / 2010.

Cependant, le mode de fonctionnement des entreprises a-t-il été impacté ? Si on exclut les entreprises *digital native*, dont toute ou partie de l'activité tourne autour de ces sujets, la réponse est pour le moment pour la plupart négative. Les algorithmes sont soit créés par des *data scientist*s plus ou moins isolés, soit achetés sur étagère par d'autres entreprises, et une grosse partie des personnes travaillant au sein de l'entreprise n'a que deux options : rester sceptiques et continuer avec les méthodes pré-IA, ou se fier presque aveuglément aux solutions que les nouveaux modèles proposent.

L'INTERPRÉTABILITÉ : DÉCLENCHEUR DE LA RÉVOLUTION IA

Voici pourquoi nous pensons que l'interprétabilité aura un rôle de déclencheur, capable d'étendre le public capable de (et tenu à) faire des choix structurants par rapport à l'utilisation des algorithmes d'intelligence artificielle.

De plus, l'interprétabilité va permettre aux organisations de mieux assurer un transfert de connaissance, le but étant de permettre aux équipes métiers de mieux résoudre d'éventuels problèmes de production mais également d'anticiper la création de nouvelles offres, la personnalisation d'un service, la commercialisation de la donnée… Dans un service de radiologie, les modèles vont permettre aux radiologues d'affiner leur diagnostic et de réduire le délai d'attente en cas d'urgence ; dans l'industrie, la modélisation va permettre l'optimisation des plans de maintenance et ainsi réduire les coûts. Ce ne sont que deux exemples parmi d'autres possibles de secteurs dans lesquels c'est l'interprétabilité qui permet une utilisation sereine des modèles.

La place de l'interprétabilité dans les questionnements d'aujourd'hui est de plus en plus orientée vers une volonté d'aide à la décision : le succès d'une

telle démarche dépendra du nombre de personnes qui seront capables de « se faire aider » par les algorithmes. Ce positionnement du « déclencheur interprétabilité » rassure également quant aux potentielles critiques : grâce à cette notion, on trouve un regain de sens et de responsabilisation des équipes. Dans de nombreuses situations ce n'est pas aux algorithmes de décider (ce qui impliquerait un flou de responsabilité) mais bien à l'humain qui dispose d'un outil d'aide à la décision.

Est-ce qu'une plus vaste interprétabilité des modèles représentera pour l'IA un événement déclencheur crucial, comme le taylorisme le fut lors de la révolution industrielle ? Est-ce qu'on commencera à intégrer des cours d'interprétation des résultats d'IA à l'école primaire, comme fut le cas pour les sujets informatiques ? Il est trop tôt pour le dire, et nous ne ferons pas semblant dans ce livre de prévoir l'avenir (surtout en l'absence d'au moins un réseau de neurones bien entraîné !). Les questions que se posent (et nous posent !) les grandes entreprises qui entament leur transformation data sont des signes de l'importance qu'une telle démocratisation du fonctionnement des algorithmes d'analyse avancée des données aura dans les prochaines années, et il serait risqué et myope de ne pas les prendre en compte en temps utile.

Est-ce qu'une plus vaste interprétabilité des modèles représentera pour l'IA un événement déclencheur crucial, comme le taylorisme le fut lors de la révolution industrielle ? Est-ce qu'on commencera à intégrer des cours d'interprétation des résultats d'IA à l'école primaire, comme fut le cas pour les sujets informatiques ? Il est trop tôt pour le dire, et nous ne ferons pas semblant dans ce livre de prévoir l'avenir (surtout en l'absence d'au moins un réseau de neurones bien entraîné !). Les questions que se posent (et nous posent !) les grandes entreprises qui entament leur transformation data sont des signes de l'importance qu'une telle démocratisation du fonctionnement des algorithmes d'analyse avancée des données aura dans les prochaines années, et il serait risqué et myope de ne pas les prendre en compte en temps utile.

MAYA AZOURI

B. L'IMPACT SUR LA PRISE DE DÉCISION

La question de l'intégration des modèles issus de la *data science* au sein des organisations est loin d'être triviale. En raison d'une nature fréquemment « boîte noire » des approches utilisées, **on peut se trouver dans la situation paradoxale d'avoir gagné en pouvoir prédictif mais sans pouvoir en tirer de leviers d'action, en raison d'une moindre interprétabilité**. Bien que des modèles prédictifs étaient déjà utilisés auparavant, donc, la démarche data science s'en différencie et présente des enjeux d'intelligibilité qui lui sont inhérents.

MODÈLES STATISTIQUES ET PRISE DE DÉCISION : DES CAS D'USAGE DÉJÀ ÉPROUVÉS

La prise de décision basée sur l'exploitation de données et modèles statistiques (études, tableaux de bord, prédictions, simulations) existe depuis longtemps dans plusieurs contextes et avec plusieurs objectifs. **On peut penser tout d'abord à la sphère publique, où des administrations telles que le Ministère de l'Économie ou l'INSEE fournissent régulièrement des prévisions de nombreux indicateurs économiques**, tels que le Produit Intérieur Brut. Les évaluations d'impact sont également usuelles : quel serait, par exemple, l'impact d'une baisse de charges patronales sur le taux d'investissement des entreprises ?
Les modèles statistiques sont développés au sein du secteur privé aussi. On peut notamment penser au secteur banque-assurance (analyse de risque ou analyse de primes d'assurance), ou encore aux nombreuses études statistiques en santé dans le cadre d'essais cliniques et développement de nouveaux médicaments. Pour finir, l'analyse statistique, plutôt descriptive, est largement utilisée au sein des services marketing, dans une logique de reporting orienté métier (la Business Intelligence).

Dans ces contextes, des modèles statistiques explicables sont utilisés. On retrouve par exemple des modèles économétriques traitant de données dites « de panel » (concernant plusieurs individus observés à différents moments du temps) et des modèles de survie (permettant l'analyse de données individuelles non observées sur une durée d'existence complète). Ces modélisations sont maîtrisées par une réflexion théorique approfondie sur le phénomène.

A l'aide de coefficients ou de règles logiques, on peut analyser de façon transparente le fonctionnement du modèle, tester ses hypothèses, étudier la sensi-

bilité de ses paramètres.
Dans certains cas, notamment d'essai clinique ou de mesure d'impact de politique publique, cette intelligibilité par nature est rendue nécessaire par l'objectif poursuivi : obtenir une analyse d'impact causal d'une variable *X* sur un phénomène *Y*. Dès lors qu'on souhaite analyser la magnitude de cet impact, des analyses statistiques approfondies sont nécessaires, de nombreux biais étant possibles : effets d'interactions entre variables, non représentativité des données…

LA *DATA SCIENCE* COMME NOUVEAU PARADIGME ?

Ces dernières années, la quantité de données disponibles, et en particulier les données dites non-structurées (capteurs, textuelles, images…), a explosé. Ce phénomène, couplé à une baisse drastique des coûts de stockage et de calcul, a facilité l'émergence de cas d'application de l'intelligence artificielle dans de nombreux secteurs et métiers : marketing, lutte contre la fraude, maintenance prévisionnelle…

En parallèle, la recherche autour de modèles et algorithmes d'apprentissage automatiques et profonds est prolifique. De nombreuses méthodes actuellement utilisés, tels que l'algorithme XGBoost, sont issues de papiers de recherche récents[6] mais ils sont pénalisés par **leur manque de transparence et d'intelligibilité : contrairement aux modèles utilisés traditionnellement, il est impossible de caractériser précisément l'impact d'une variable d'entrée sur la variable de sortie à partir d'une simple lecture de la modélisation.** Il faut des processus ex-post qui, à partir par exemple d'analyses de sensibilité, permettent de quantifier a *minima* des contributions.

Les modèles d'apprentissage automatique ont pour objectif premier d'apprendre des données, prévoir des valeurs, sans avoir le souci de caractériser précisément et de façon séparable le lien entre les variables d'entrée et la variable d'intérêt. **Ces nouveaux modèles ne remplaceront donc pas les modèles économétriques dans certains cas, notamment pour les analyses causales : chaque méthodologie a son intérêt propre.**

[6] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

LES NOUVEAUX QUESTIONNEMENTS PROPRES À LA DATA SCIENCE

Les avancées théoriques par rapport à l'interprétabilité des modèles permettent un certain optimisme quant à leur avenir en entreprise. En revanche, ils ne dispensent pas d'avoir une réflexion approfondie sur le rôle de la data et des modèles au sein du processus de prise de décision. Tout d'abord, souhaite-t-on un modèle qui prend seul une décision, ou qui aide un analyste à la prendre ? Quel mécanisme de partage de responsabilité entre l'analyste et le modèle envisage-t-on, notamment en cas de « mauvaise » décision ?
Supposons que le modèle ait été rendu plus transparent, le sujet de la conduite du changement demeure. On peut développer un modèle parfait d'un point de vue mathématique, si l'analyste ne l'utilise pas (pour diverses raisons, fonctionnalité difficile à accéder par exemple), une meilleure compréhension ne changera pas forcément cet état de fait.
Par ailleurs, en tant que décideur, il est important d'avoir conscience de plusieurs points :

- **Ce que le modèle intègre et n'intègre pas.** Par exemple, les classements d'universités américaines intègrent la production académique et le pourcentage d'étudiants étrangers, mais pas le coût d'inscription : si ce dernier était inclus, le classement serait totalement différent. Les modèles reflètent donc une certaine réalité, propre à sa conception.

- **La nécessité d'avoir un regard critique sur les éventuels biais du modèle** (on peut notamment penser à l'étude du MIT qui montre que plus la carnation d'un individu est foncée, plus la reconnaissance faciale voit son taux d'erreur augmenter, jusqu'à 35% pour les femmes noires). Ce type de sujets doit être pris en compte dès le design du modèle, avec par exemple des « bonnes pratiques » d'entreprise ou une charte éthique et méthodologique. D'importants chantiers concernant l'auditabilité des modèles vont s'ouvrir, notamment en santé.

- **Les changements générés sur la prise de risque par l'intégration d'un outil d'aide à la décision** : si en cas d'erreur éventuelle la faute tombera sur l'IA si l'utilisateur a confirmé sa suggestion, et sur lui-même s'il a donné un avis discordant, on pourra s'attendre une diminution rapide du rôle de contrôle des experts métier.

La prise de décision sera ainsi plus ou moins automatisée selon le contexte, à mesure que la data deviendra de plus en plus centrale au sein des organisations.

L'intégration de modèles au sein des processus pose la question du niveau de responsabilité entre analyste et algorithme, le niveau d'automatisation et d'intelligibilité souhaité, la bonne utilisation par les analystes. Par ailleurs, les nouveaux modèles ne viennent pas en remplacement des approches déjà en œuvre, notamment sur l'évaluation des politiques publiques, mais plutôt en complément.

AURÉLIA NÈGRE

ILS ONT TENTÉ L'EXPÉRIENCE

GRTgaz

Les prédictions des niveaux de stock ne seront jamais totalement automatisées et envoyées telles quelles, sans validation humaine. Il y a des informations logistiques (renouvellement d'un marché, références qui vont être mises en fin de vie...) qui ne sont pas prises en compte à date par le modèle. En revanche certains consommables à faible enjeu et à consommation régulière pourraient avoir des niveaux de stock totalement automatisés à terme. Dans le futur, une GMAO connectée permettra d'anticiper les pannes et d'en déduire des niveaux de stock optimaux. Ce sera bien la complémentarité entre l'homme et la machine qui permettra d'aller plus loin.

PSA

Sans interprétabilité, il est plus difficile d'aller en production. Le modèle de *machine learning* offre une vision probabiliste qui peut perturber certains métiers qui ont pour habitude des chiffres exacts. Il faut les rassurer sur le fonctionnement et les limites du modèle, et démontrer qu'on fait mieux qu'avant en termes de gain. Les modèles développés aident à préparer la prise de décision. Le modèle doit donc soumettre ses résultats de manière la plus justifiée possible puisque ce n'est pas le modèle qui prend la décision à la place de l'homme.
En revanche, pour les « micro décisions » quotidiennes qui ont un niveau de complexité faible et un impact minime en cas d'erreur, il serait possible de déléguer totalement la prise de décision à la machine : par exemple pour un moteur de recommandation ou la personnalisation des pages web.

C. CRÉER LES CONDITIONS POUR UNE RÉVOLUTION PÉRENNE

Les nouveaux questionnements posés par la *data science* et notamment l'éventualité d'utiliser des modèles « black box » remettent en cause les structures organisationnelles classiques de l'entreprise (découpage fonctionnel, divisionnel ou matriciel) et une organisation centrée autour de la donnée (*data-driven*) devient une nécessité face à la concurrence des pays à faible coût de main d'œuvre et aux géants du web. **Les principales caractéristiques d'une entreprise *data-driven* peuvent se résumer à la facilité :**

- **De collecter des données** fiables et pertinentes provenant de différentes sources ;
- **D'exploiter son patrimoine de données** pour disrupter ses processus et créer des offres de service innovants.

La création de nouveaux services grâce à la donnée ne peut se faire qu'en intégrant de nouveaux process et nouveaux rôles au sein de l'entreprise, afin de garantir la conformité légale d'une part, et l'alignement entre l'innovation et la mission de l'entreprise d'autre part. La gouvernance de la donnée permet de garantir une compréhension partagée par tous des décisions prises grâce à la donnée récoltée, et sera donc la garantie d'une transparence accrue, et d'une prise de décision optimale. Dans ce nouveau modèle d'organisation, chaque nouveau rôle détient une part de responsabilité : de la récupération d'une donnée non biaisée à l'activation opérationnelle d'une décision, en passant par la compréhension technique ou métier d'un modèle.

DE NOUVEAUX RÔLES OPÉRATIONNELS

Les *data scientists* jouent un rôle clé dans cette organisation en construisant **des modèles analytiques** (fondés sur des algorithmes de *machine learning* ou de *deep learning*[7] par exemple) **capables de transformer de grands volumes**

[7] L'apprentissage profond (***deep learning*** en anglais) est un ensemble d'algorithmes utilisés dans l'apprentissage machine, utilisés pour modéliser des abstractions de haut niveau dans les données à l'aide d'architectures de modèles, qui se composent de multiples transformations non linéaires.

de données en nouvelles informations qui permettent de faciliter la prise de décision. Pour l'instant, elle est réalisée par des opérateurs humains qui peuvent se demander comment faire confiance à ces modèles. Par exemple, une information contre-intuitive issue d'un algorithme va générer davantage de questions que de réponses. Si les *data scientist*s peuvent lister les données utilisées en entrée mais ne sont pas en mesure d'expliquer les résultats ni le raisonnement derrière ces algorithmes (réseau de neurones par exemple), ces modèles ont une forte probabilité de rester dans un tiroir. **En complément des *data scientist*s, les *data engineer*s[8] deviennent indispensables pour industrialiser l'environnement capable d'accueillir les modèles créés.** En partant de ce constat, la clé pour exploiter pleinement la data réside dans la capacité à expliquer une décision prise grâce à une maîtrise de ses données, de l'interprétation des modèles et des résultats en sortie.

NÉCESSITÉ D'UNE GOUVERNANCE EN SUPPORT DES PROJETS DATA

La gouvernance de la donnée correspond à un ensemble de rôles, de processus et de responsabilités dans une organisation, doublé d'une vision et d'une culture qui ont pour objectif de démocratiser l'accès, l'exploitabilité à une donnée définie, de qualité, de façon sécurisée et éthique.

Le levier principal pour répondre à ces objectifs repose sur un bon leadership pour créer une culture qui va mettre les données au cœur de l'organisation. La mise en place de cette doctrine va favoriser la compréhension et la conviction des employés dans l'importance des données dans la prise de décisions. Les entreprises ont rapidement reconnu que le succès de la data science ne repose pas que sur les *data scientist*s mais **exige de nouveaux rôles qui soient capables d'interpréter les modèles, de les traduire et de s'assurer que les organisations obtiennent un impact réel de leurs initiatives analytiques.**

Ces personnes capables d'interpréter les résultats, le fonctionnement des modèles et de les traduire ne sont pas obligatoirement des experts de la donnée. Par exemple, le *deep learning* permet d'assister les professionnels de la santé

[8] Un *data engineer* est un travailleur dont les principales responsabilités professionnelles consistent à préparer des données à des fins analytiques ou opérationnelles. Ses tâches spécifiques comprennent généralement la construction de pipelines de données pour rassembler des informations provenant de différents systèmes sources, l'intégration, la consolidation et le nettoyage des données, et leur structuration pour utilisation dans des applications analytiques individuelles.

dans l'analyse d'imagerie médicale. Ces personnes (non expertes en data science) doivent être capables de comprendre et d'interpréter les résultats issus de ces analyses pour éviter les erreurs de diagnostic. Elles jouent un rôle essentiel dans une organisation *data-driven* en faisant le lien entre ces experts techniques et les métiers (marketing, supply chain, operations, risques…).

Ces nouveaux rôles présentent la particularité de se répartir dans l'ensemble de l'organisation côté métiers et côté IT avec des capacités d'interprétation et de traduction différentes :

- Maîtrise de la donnée (où est-elle stockée ? Comment est-elle utilisée (*data lineage*) ? Qui y a accès ? Comment est-elle ingérée ? À quelle fréquence est-elle mise à jour ? Est-ce une donnée sensible ?) ;
- Maîtrise de l'interprétabilité des modèles, capacité à expliquer le raisonnement derrière un algorithme ;
- Maîtrise de l'interprétabilité des résultats, capacité à expliquer les résultats issus d'un modèle créé par un *data scientist* ;
- Maîtrise d'une utilisation éthique de la donnée, en conformité avec les valeurs et la mission de l'entreprise.

Toute personne concernée par la démarche de l'exploitation de la donnée s'appuie sur des processus (comment fiabiliser les données ? comment partager les données ?) **pour mettre en œuvre des règles et des normes** (quel niveau de qualité ? quel niveau d'explicabilité et d'interprétabilité d'un modèle ?) qui garantissent des données gouvernées au cœur de l'organisation. Par exemple, imaginons un *data scientist* qui se lance sur un projet d'attrition en détectant les clients insatisfaits d'une banque et qui risquent de la quitter. Pour cela, il s'appuie sur les données client et l'historiques des échanges (emails, appels, rendez-vous avec le conseiller…). La réussite de ce projet commence par le respect des contraintes réglementaires (notamment du RGPD pour toute utilisation des données personnelles) qui est responsabilité du DPO, et par la vérification de la fiabilité des informations contenues dans les bases de données, responsabilité du *data steward*. Le *data owner* s'assure que les données en entrée et l'information créée par le modèle aient une interprétation réelle du point de vue métier. Enfin, les comités d'éthique que nous voyons apparaître s'assurent que l'utilisation de la donnée et la prise de décision qui en résulte restent en phase avec la politique de « bonne conduite » de l'entreprise.

RÔLE	MAÎTRISE DE LA DONNÉE	MAÎTRISE DE L'INTERPRÉTABILITÉ DES MODÈLES	MAÎTRISE DE L'INTERPRÉTABILITÉ DES RÉSULTATS
CHIEF DATA OFFICER • Diffuser une culture orientée autour de la donnée • Garantir la définition et la mise en place d'une gouvernance de la donnée à l'échelle de l'entreprise	●●	●●	●●●●●
DATA PROTECTION OFFICER • Assurer la protection des données personnelles • Favoriser la création d'une culture de transparence et de conformité en matière de traitement des données	●●	●	●●●
DATA OWNER • Réfléchir à des nouveaux cas d'usage autour de la donnée • Délivrer les autorisations d'accès • Assurer la maturité des données	●●●	●	●●●
DATA MANAGER • Définir et conduire la stratégie de gouvernance des données (avec le CDO) • Assurer que les projets et les cas d'utilisation appliquent les règles de la gouvernance des données • Faciliter la mise en œuvre des besoins des Data Owner • Superviser la mise en œuvre des règles de gouvernance par les Data Steward	●●●●	●●	●●
DATA STEWARD • Mettre en œuvre les règles de gouvernance des données • Assurer l'accessibilité demandé et la sécurité des données	●●●●●	●●	●
DATA SCIENTIST • Construire des modèles qui vont valoriser les données	●●●	●●●●●	●●●●
DATA ENGINEER • Concevoir l'environnement qui va accueillir les modèles des Data Scientist	●●●●	●●●●	●●
UTILISATEURS DE DONNÉES • Exploiter les données disponibles • Être en capacité d'expliquer les informations issues des modèles créées	●	●	●●●●●

Légende : Plus le nombre d'● est important, plus la maîtrise est importante

DE NOUVEAUX RÔLES, MAIS AUSSI DE NOUVEAUX PROCESSUS

Pour s'assurer que la donnée reste bien au centre de l'organisation, une culture d'explicabilité et d'interprétabilité doit être généralisée à tout niveau de l'entreprise. Le *Chief Data Officer* garantit ce partage et cette diffusion de la connaissance en mettant en place une approche transversale au sein des équipes. En s'appuyant sur la donnée gouvernée comme moyen de fédérer les personnes, l'attractivité de collecter des nouvelles sources et d'exploiter un patrimoine élargi va créer un cercle vertueux qui intégrera les enjeux d'interprétabilité.

JULIEN VONG

ILS ONT TENTÉ L'EXPÉRIENCE

GRTgaz

La construction du modèle s'est faite en binôme entre les utilisateurs métier et l'équipe Data Science. Cette co-construction et de nombreux tests sont nécessaires pour la compréhension des résultats, notamment le calcul des écarts entre les anciennes prédictions et les nouvelles prédictions pour chaque pièce. En cas de dérive du modèle, le Datalab sera là en support.
L'extraction des données pour alimenter le modèle et l'interprétation des résultats d'un point de vue métier reposent aujourd'hui sur une seule personne, ce qui représente un risque potentiel de perte d'information. Si le modèle doit être révisé suite à un changement de stratégie sur les stocks et donc sur les règles métiers, cela impliquera une mise à jour des variables en entrée et du code par la DSI.

PSA

Un projet ne démarre pas tant qu'il n'y a pas eu des échanges réguliers entre l'équipe Data Science et le métier pour dialoguer sur les futurs résultats du modèle et la nature de la décision. Lors de l'expérimentation, on estime que la charge de développement informatique est de 60% du délai alors que la charge liée à la mise dans la main du métier, l'apprentissage, l'optimisation et la transformation des processus représente 40% du délai. Ce ratio s'inverse à la phase pilote afin de permettre une plus grande appropriation par le métier.

2. RÈGLES ET SOCIÉTÉ

Les outils et services basés sur l'intelligence artificielle se développent dans une multitude de secteurs et d'applications. Un nombre croissant de décisions sont ainsi prises par des modèles prédictifs, ayant exploités les corrélations existantes dans une base de données à disposition. La question de la responsabilité vis à vis des décisions prises pourra se décliner plutôt en termes de principes éthiques (ce qui est souhaitable) ou d'exigences réglementaires (ce qui est légal), mais sera dans tous les cas au cœur du sujet. Plus précisément, l'intelligence artificielle ne procédant pas par intention elle n'est pas responsable en tant que telle et la répartition de cette responsabilité doit être précisée par des textes et des normes. **La première partie du chapitre aborde ainsi l'intelligibilité au sein de la réglementation actuelle et l'appui qu'elle peut apporter au concept de redevabilité de l'IA.**

Chaque secteur d'activité, institution, individu possède des visions et des objectifs extrêmement diversifiés vis à vis de l'IA. Pour de nombreux cas d'usage l'écosystème est complexe et les parties prenantes sont non seulement multiples, mais avec des intérêts potentiellement non compatibles. **C'est pourquoi dans un deuxième temps une réflexion autour des attentes des différents acteurs sociaux vis à vis de l'interprétabilité des algorithmes sera proposée.**

Enfin, du point de vue des protagonistes de l'IA, l'arrivée de nouvelles contraintes réglementaires peut laisser planer le doute d'un ralentissement de l'activité et des opportunités, surtout en Europe. Si le nombre de problèmes qu'elles posent dans les applications actuelles de l'IA sont important, il reste à déterminer s'ils sont de « bons problèmes » initiateurs d'un changement nécessaire et inévitable des pratiques **: dans un troisième temps les conséquences de ces contraintes et leur impact sur l'innovation seront examinés**


PIC DU MODÈLE EN PROD
ÇA VA ÊTRE LONG POUR Y ARRIVER!
TÉLÉSIÈGE DE L'INTERPRETABILITÉ
PAS VRAIMENT!

RÉGLEMENTATION ET EXIGENCES D'EXPLICATION POUR LES ALGORITHMES

Comme nous vivons dans une période de bouleversements majeurs sur les sujets data, on pourrait avoir tendance à oublier que la réglementation n'a pas attendu l'essor du Big Data et l'IA pour construire des textes qui encadrent l'usage d'algorithmes. Au contraire : en France, par exemple, la loi Informatique et Libertés de 1978 abordait déjà ces thématiques. La dernière avancée réglementaire, capable de mettre au cœur du débat public le sujet, est le RGPD : cette réglementation générale en vigueur depuis mai 2018, insiste surtout sur la nature des données sources (données personnelles de ressortissants européens) et moins sur le type de décisions prises. Afin de rendre ce cadre encore plus complexe, à ces directives globales s'ajoutent des textes spécifiques à secteurs d'activité sensibles comme la banque, qui permettent de les réguler de plus près...

Qu'est-ce que tous ces textes partagent donc ? De manière générale, ils imposent de donner des explications sur l'algorithme utilisé, voire sur la manière dont il a été mis au point : mais ces exigences ne sont pas de même nature ni de même degré dans les différents cas de figure.
Ici, nous n'avons pas l'ambition de couvrir tous les aspects des différentes réglementations : plutôt, nous insisterons sur le type d'explications demandées du point de vue des traitements réalisés sur les données : certaines exigences, en particulier, peuvent être difficiles à satisfaire pour défaut d'intelligibilité de certains modèles prédictifs complexes largement utilisés dans la pratique. Ce manque à donc deux impacts : il complexifie le respect et l'interprétation de certaines réglementations à court terme, et il rend difficile la formalisation de textes réglementaires portant sur les exigences demandées aux modèles prédictifs à moyen-long terme.

LA RÉGLEMENTATION GÉNÉRALE

Concernant les exigences portant sur la manière dont sont prises des décisions, le Règlement Général sur la Protection des Données (RGPD) relaye trois principes de la loi Informatique et Libertés de 1978, en précisant :

- La nécessité de bien identifier certains éléments du traitement automatisé

de données personnelles, en particulier ses données sources et sa finalité ;

› L'interdiction qu'une machine puisse prendre une décision seule (sans intervention humaine) entraînant des conséquences cruciales pour une personne (décision judiciaire, décision d'octroi de crédit...) Elle ne s'applique pas en cas de consentement explicite, ou lorsque le traitement est nécessaire à l'exécution du contrat ;

› Le droit pour les personnes d'obtenir des informations sur la logique sous-jacente du fonctionnement de l'algorithme

Le premier niveau d'exigence est ainsi équivalent à communiquer sur les sources de données personnelles utilisées dans les traitements. Les challenges qu'il entraîne sont principalement organisationnels et liés à la communication externe, mais ne posent pas de difficultés particulières dues à la complexité des algorithmes utilisés.
Les implications du dernier point, au contraire, peuvent varier énormément selon la manière dans laquelle il est interprété. S'agit-il de davantage détailler les variables explicatives les plus importantes ? Ou bien de communiquer sur la ou les grandeurs qui sont optimisées par l'algorithme ? Ou encore, d'aller plus loin en détaillant comment se comportent les variables importantes, et s'il y a par exemple des effets importants sur les prédictions une fois passés certains seuils ? Aujourd'hui, l'absence de jurisprudence ne permet pas d'avoir des certitudes : en cas de traitement complexe, dans tous les cas, formuler la logique sous-jacente du traitement et la rendre intelligible n'est pas tâche aisée.

Par ailleurs, certains angles morts persistent dans la réglementation actuelle. Les articles du RGPD ne s'appliquent en effet qu'aux données personnelles exclusivement, alors que certains traitements automatisés ont potentiellement des effets collectifs majeurs (pensons à la répartition d'un budget entre différents établissements ou communes, au plan de déploiement d'une nouvelle technologie...) sans pour autant s'effectuer à l'échelle individuelle : ceux-ci ne sont donc pas concernés par le RGPD.

UN EXEMPLE DE RÉGLEMENTATION SECTORIELLE

Dans le secteur bancaire, les organismes doivent calculer leur exposition au risque de crédit et avoir à disposition des fonds propres, dimensionnés pour cou-

vrir un certain nombre de scenarii dans le cas de défauts sur des crédits. Depuis 2007, la norme Bâle II qui a été retranscrite en directive européenne et qui est en vigueur pour la zone euro précise que, par défaut, c'est la méthode standard qui s'applique pour calculer les fonds propres, en s'appuyant sur des systèmes de notations d'organismes externes.

La norme laisse néanmoins la possibilité aux banques d'utiliser leurs propres modèles internes de calcul de risque. Dans ce cas, les équipes de modélisation ont des obligations précises (et récemment réaffirmées dans le guide TRIM de la Banque Centrale Européenne) dont en voici une partie :

- Disposer d'un historique de 5 ans minimum avec preuve de complétude, qualité et représentativité ;
- Justifier que les données externes sont bien utiles et suffisamment fiables pour être utilisées en production - il y a un risque exogène de voir disparaître ou changer le flux de données en question ;
- Structurer les différents modèles afin qu'ils prédisent différents ratio imposés (probabilité de défaut, exposition au moment du défaut, taux de perte en cas de défaut) ;
- Justifier les choix de sélection de variables, de discrétisation de variables, de croisement de variables...
- Effectuer un suivi régulier de la performance et de la stabilité (*backtesting*) au moins annuel en fixant à l'avance des seuils quantitatifs de « point d'attention » et « alerte », sur des indicateurs à surveiller.

Afin d'être en conformité avec ces principes, une organisation particulière est souvent adoptée en séparant les **équipes modélisation** qui créent les modèles prédictifs et justifient les partis-pris de modélisation et **les équipes validation** qui analysent le modèle et font des préconisations priorisées sur les éléments à consolider voire modifier.
Au-delà de l'organisation et de la gouvernance, un point essentiel est que les obligations vont bien plus loin qu'exiger seulement des preuves de performance. Elles s'étendent ainsi à des domaines incluant la qualité de donnée en production, le suivi dans le temps du comportement du modèle et la justification des étapes de modélisation (en mettant à disposition du régulateur des éléments précis et en les justifiant).

Ce dernier point est déjà complexe lorsqu'on utilise des modèles souvent présentés comme « simples », qui réalisent des sommes pondérées de variables explicatives discrétisées (régression logistique). Mais le caractère peu intelligible des modèles plus complexes rend encore plus difficile la formalisation d'un texte de loi concernant les exigences autour de leur explication. Il est ainsi nécessaire que les acteurs, modélisateurs comme régulateurs, continuent à innover sur la forme et le minimum requis du contenu de ces justifications.

COMPRENDRE LES MODÈLES, COMPRENDRE LES RÉSULTATS : UN DÉBAT OUVERT

Face à la complexité de certains algorithmes, plusieurs approches sont possibles concernant le cadre à donner à la réglementation.

Tout d'abord, la question se pose de savoir si les exigences devraient portent **sur la manière dont l'algorithme a été créé** et entraîné avec les données, ou alors si les **exigences concernent plutôt l'algorithme créé**, qui devrait alors valider un certain nombre de propriétés. Dans le premier cas, on cherche à intégrer un maximum de propriétés à satisfaire dès la construction et de communiquer avec le régulateur sur ces propriétés. On parle de *responsibility by design*, et un exemple simple serait d'imposer une dépendance connue entre une entrée et la sortie prédite (par exemple, si la surface du logement augmente, alors le prix prédit ne peux pas diminuer). Dans le second cas, on cherche à prouver que certaines propriétés sont vérifiées dans le maximum de cas possibles, mais on raisonne seulement à partir de l'algorithme sans s'intéresser à la manière dont il a été construit. On cherche ainsi une validation a posteriori, qui doit se baser sur une méthodologie et des outils adaptés.

L'exemple détaillé précédemment sur la réglementation en risque de crédit impose des exigences à la fois sur la manière dont l'algorithme a été créé (justification des étapes de sélection de variables, croisement de variables…) et sur les propriétés de l'algorithme obtenu (les coefficients associés à chaques variables). Cependant, dans d'autres contextes et secteurs, on peut tout à fait imaginer que l'exigence ne porte principalement que sur un de ces deux aspects. Selon l'axe choisi, les questions qui se posent pour la recherche sont différentes :

- Comment développer des méthodologies pour construire un modèle qui va imposer et valider certaines dépendances connues ou souhaitées entre entrées et sorties ?

› Comment mettre à disposition des outils permettant d'analyser, d'explorer et de donner du sens à tous les paramètres internes d'un modèle prédictif, afin de mieux comprendre comment le modèle réalise ses prédictions ?

Les exigences réglementaires et les bonnes pratiques qui se diffusent dépendront en partie des avancées techniques sur le sujet de l'intelligibilité du *machine learning*, ainsi que du contexte et de la portée des décisions prises. Il est probable que le sujet reste ouvert pour de nombreuses années et que seules des jurisprudences permettront de mieux transposer les exigences formulées à haut-niveau en langage précis pour les équipes de modélisation data.

JEAN-MATTHIEU SCHERTZER

ILS ONT TENTÉ L'EXPÉRIENCE

BPCE

A BPCE, l'équipe Modèles internes & Big Data de la DRCCP porte la responsabilité de la performance et de l'interprétabilité des modèles devant le régulateur (ACPR, BCE) et doit justifier les choix faits sur les modèles. Cette responsabilité impose de respecter un certain nombre de procédures et de bonnes pratiques tout au long des différentes étapes de mwodélisation : i. choix de variables, ii. sélection de variables et iii. choix et entraînement des modèles.
Le choix des variables explicatives repose sur des discussions avec des experts afin de s'assurer que les variables injectées dans le modèle correspondent à des ratios ou agrégats qui ont un sens métier. En effet, l'intelligibilité du modèle passe en premier lieu par celle des variables.

La sélection de variables consiste à retenir les variables qui contribuent à « bien prédire » les défauts de crédit. Une première étape consiste à effectuer cette sélection en optimisant un critère statistique. C'est à ce titre que des modèles peu interprétables (random forest par exemple) peuvent être utilisés. Ainsi listées, les variables à garder sont examinées par des experts afin de valider la pertinence de leur présence. Ici, le caractère peu interprétable des modèles

sont entraînés n'est pas bloquant car les décisions prises sont validées manuellement.

Le choix et l'entraînement des modèles de crédit est soumis à des exigences réglementaires, qui ont pour objet une intelligibilité très forte du comportement du modèle. En pratique, pour pouvoir être homologués, seuls certains types de modèles réputés interprétables (régression logistique, arbre de décision, modèle de survie, etc.) peuvent être choisis. Obtenir une performance satisfaisante tout en utilisant ces modèles interprétables est bien sûr possible mais a un coût : il est nécessaire d'effectuer un travail fin sur les variables (discrétisation, croisement, etc.), là où certains modèles plus complexes sont plus performants sur les variables brutes.

Du côté des perspectives, certains tests sont menés avec du *Deep Learning* (réseaux de neurones profonds) pour modéliser le risque de crédit. L'objectif est de tester et mesurer la valeur ajoutée de ces modèles, sans les mettre en production aujourd'hui. L'analyse des dossiers de crédit pour lesquels les prédictions sont différentes de celles des modèles en production permettra de mieux mesurer leur performance et leur robustesse. De telles analyses permettront de continuer à ouvrir la discussion avec le régulateur, concernant le domaine de validité de ces modèles moins intelligibles et plus complexes.

CHUS

Il faut une éthique des algorithmes publics auto-apprenants, qui doivent servir des objectifs louables tout en évitant à long terme des changements insidieux dans les résultats. C'est le rôle des organismes publics de réglementer la vérification régulière et de s'assurer de l'absence de biais dans les algorithmes. Les data scientists qui travaillent sur des algorithmes liés à la santé pourraient s'engager formellement à respecter des bonnes pratiques et un code éthique, de manière similaire aux médecins qui s'engagent à respecter le Serment d'Hippocrate. Une autre proposition serait la mise en place d'une agence d'audit détachée de l'autorité administrative qui pourrait statuer sur les traitements algorithmiques pour lesquels une présomption de non-conformité aux principes éthiques aurait été enregistrée.

PSA

D'un point de vue réglementaire, le RGPD ne limite pas l'interprétabilité ni la performance des modèles chez PSA. Par exemple, les données clients anonymisées sont suffisantes pour les usages statistiques prévus du véhicule connecté comme les flux de circulation quotidiens.

B. DES ATTENTES HÉTÉROGÈNES POUR DES ACTEURS SOCIAUX VARIÉS

L'utilisation de plus en plus fréquente d'algorithmes d'intelligence artificielle fait naître plusieurs besoins, parfois contradictoires. En effet, de nombreux rôles interviennent dans les différents moments du cycle de vie de l'algorithme. En s'intéressant aux interactions entre eux, nous pourrons mieux identifier et décrire leurs attentes.

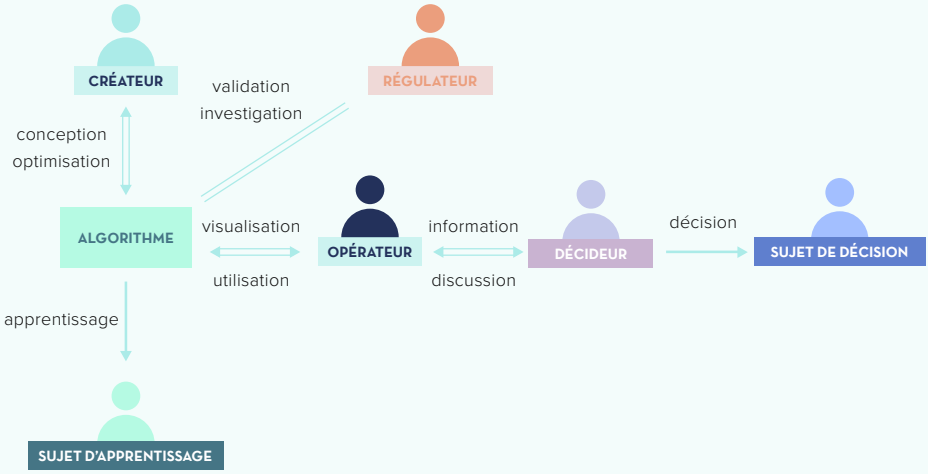


Figure 1 : Environnement d'un algorithme d'intelligence artificielle. Les flèches indiquent le sens des interactions. Chaque acteur est animé par des considérations propres, et il a des attentes particulières vis-à-vis de l'interprétabilité.

En général, on peut identifier six rôles principaux qui composent l'environnement d'un algorithme d'intelligence artificielle :

› **Le créateur,** qui a développé et mis en production l'algorithme, ou en détient la propriété intellectuelle. Dans la plupart des situations il existe un multiplicité de créateurs ayant plusieurs profils (le *data scientist* qui a réalisé le *POC*[9] le *data engineer* qui l'a rendu opérationnel, etc.), voire même étant

[9] Un POC, ou *Proof of Concept*, est un un « avant prototype » simplifié au maximum d'un projet, qui permet d'en tester l'intérêt concrètement et avec un effort limité.

de plusieurs entités différentes (le créateur de la brique algorithmique réutilisable, open source ou non) ;

› **L'opérateur,** qui interagit directement avec l'algorithme, en renseignant les données et en collectant les résultats ;

› **Le décideur,** souvent plus haut dans la hiérarchie, qui prend la décision finale sur la base d'une synthèse des résultats fournie par l'opérateur ;

› **Le sujet de décision,** qui est directement affecté par la décision assistée par intelligence artificielle ;

› **Le sujet d'apprentissage,** dont les données ont servi à l'apprentissage du modèle d'IA ;

› **Le régulateur,** qui est en charge d'évaluer le modèle et de faire respecter la loi.

Chacun de ces rôles s'accompagne d'objectifs et de préoccupations propres, desquelles découlent une certaine vision de ce que représente l'interprétabilité. Ces rôles peuvent se confondre au sein d'une même entité ou au contraire se répartir sur plusieurs acteurs. Qui plus est, tous ces acteurs ne sont pas forcément des personnes, mais peuvent représenter des machines ou même des niveaux plus abstraits d'intelligence artificielle.

CAS D'USAGE	OCTROI DE CRÉDIT BANCAIRE	PRISE DE TRAITEMENT MÉDICAL	CHOIX D'ORIENTATION UNIVERSITAIRE
CRÉATEUR	BANQUE	ENTREPRISE EXTERNE	ENTREPRISE EXTERNE
OPÉRATEUR	CONSEILLER	PERSONNEL MÉDICAL	PERSONNEL UNIVERSITAIRE
DÉCIDEUR	DIRECTEUR	MÉDECIN/PATIENT	PROFESSEUR/ÉTUDIANT
SUJET DE DÉCISION	CLIENT	PATIENT	ÉTUDIANTS
SUJET D'APPRENTISSAGE	CLIENTS PRÉCÉDENTS	PATIENTS PRÉCÉDENTS	ÉTUDIANTS PRÉCÉDENTS
RÉGULATEUR	BCE*	HAS**	MINISTÈRE DE L'ÉDUCATION

*Banque Centrale Européenne

** Haute Autorité de la Santé

Les six rôles sont illustrés concrètement sur le tableau ci-dessus. A titre d'illustration, nous avons représenté trois cas d'usage distincts :

› Une banque veut décider des modalités d'octroi de crédit (montant, taux d'intérêts, ou refus) pour des particuliers, sur la base d'un algorithme calculant la probabilité que le client fasse défaut ;

› Un hôpital veut décider d'un traitement à donner à un patient sur la base d'un diagnostic et de recommandations assistés par intelligence artificielle ;

› Une plate-forme d'orientation universitaire est en charge de dispatcher les bacheliers dans les universités qui leur correspondent le mieux.

Le créateur (par exemple un *data scientist*), recherche la transparence afin **d'optimiser une métrique de performance donnée.** Il cherche également l'*explicabilité* pour identifier ou créer des variables discriminantes et porteuses de sens pour la décision à prendre. La performance portant sur un ensemble de données, c'est également **l'aspect *global* de l'interprétabilité** (c'est à dire agrégé sur plusieurs observations) qui est recherché. En s'intéressant ainsi à la manière dont sont traitées les différentes variables explicatives, le créateur cherche à réduire les biais qui seraient induits par sa modélisation du problème.

L'opérateur doit s'assurer que les données qu'il fournit à l'algorithme sont bien complètes et pertinentes pour la décision à prendre, ce qui nécessite de la t*ransparence*. Il doit également **synthétiser les résultats avant de les présenter au décideur**, et éventuellement répondre aux interrogations de ce dernier, d'où un besoin d'*explicabilité*.
Dans le cadre d'octroi de crédit, l'opérateur est le conseiller. Il intervient pendant un entretien avec le client, en remplissant en temps réel les informations que le client lui donne et en recevant les prédictions de l'algorithme sur la probabilité que ce dernier fasse défaut. En fonction de la stratégie propre à la banque, il lui communique les modalités du crédit.

Le décideur veut prendre la meilleure décision qui soit, compte tenu des contraintes applicables au problème (importance de l'objectif, budget, règlementation, délai). Il s'agit souvent de **trouver un compromis entre plusieurs attentes contradictoires** (par exemple en médecine entre un traitement lourd mais fiable ou un traitement plus léger mais incertain). L'*explicabilité* est alors l'objectif visé en priorité.

Pour un traitement médical, le rôle de décideur est partagé entre les médecins et le patient. En effet, le médecin propose plusieurs solutions au patient, mais c'est ce dernier (ou sa famille) qui *in fine* décide de la solution à envisager.

Le sujet de décision est uniquement intéressé par son propre cas, et cherche une explication à l'échelle de l'individu : **quels sont les facteurs qui ont joué en *sa* faveur ou défaveur ?** Par simple curiosité ou dans l'espoir de changer la décision prise, c'est une exigence de *contestabilité* : chacun peut vouloir être en mesure de contester la décision de l'algorithme, en identifiant quels aspects de la décision font défaut. Cela nécessite que l'algorithme soit *localement explicable*. Le sujet de décision peut également vouloir **comprendre l'algorithme pour adapter son comportement** et influer sur la décision, par exemple en modifiant l'ordre des vœux sur une plate-forme d'orientation universitaire.

Le sujet d'apprentissage est multiple : dans une application médicale, il englobe tout à la fois les patients qui ont développé la même maladie par le passé, mais également les médecins avec une expérience dans le domaine, voire les articles de publication scientifique. Dans le cas de personnes physiques, elles ne sont pas toujours conscientes que leur données aient servi à alimenter une solution d'intelligence artificielle (surtout si elles ont été utilisées de manière anonyme). **Ces utilisateurs peuvent cependant se demander, au point de vue moral, comment leur comportement peut servir à alimenter des outils d'aide à la décision pour d'autres personnes.** La réglementation s'affinant, un nombre croissant d'utilisateurs est en mesure de demander la restitution ou la suppression de leurs données personnelles, même si cette suppression peut rester sans effet sur l'algorithme en production pendant un certain temps : un réseau de neurones ne peut pas mettre à jour ses coefficients en temps réel, et ne sera donc modifié qu'à la prochaine phase de réentraînement. A contrario, un kNN[10] (*k Nearest Neighbors*) est immédiatement impacté par la suppression d'une observation car il utilise directement la donnée à sa disposition. Le sujet d'apprentissage est donc a priori intéressé par les conséquences directes que ses données ont et peut être en droit d'exiger une mise à jour rapide du modèle après la suppression de ces dernières. Il s'agit alors d'une exigence de *transparence*.

Le régulateur, quant à lui, a pour mission de réaliser un audit du système, de

[10] *L'algorithme des k voisins les plus proches (k-NN) est une méthode non paramétrique utilisée pour la classification et la régression.*

faire respecter la législation en vigueur, voire le cas échéant de procéder à des enquêtes pour identifier des dysfonctionnements. La sécurité du système peut être testée sur des données fictives ou nouvelles, tandis que les enquêtes ont besoin de connaître précisément les décisions prises dans le passé. Ce dernier point requiert le stockage de l'algorithme, de ses prédictions et de ses interprétations si elles ne peuvent pas être rejouées *a posteriori*. C'est particulièrement difficile pour les algorithmes d'apprentissage par renforcement, où le modèle se met à jour constamment en réponse à chaque nouvelle observation. **L'audit de l'algorithme se fait donc sur la base de sa *transparence*. Enfin, le régulateur peut vouloir vérifier l'*équité* de l'algorithme** en s'assurant que certains attributs jugés discriminatoires par la société ne soient pas pris en compte (même s'ils sont porteurs de sens en ce qui concerne la décision à prendre).

Il apparaît donc clairement que tous les acteurs ne poursuivent pas les mêmes objectifs d'interprétabilité, ni n'en ont une définition unique. A un niveau encore plus global, on peut également s'interroger sur **l'aspect culturel de l'interprétabilité** : l'Europe semble se détacher comme une figure de proue de la défense des intérêts des citoyens et a fait de l'interprétabilité une obligation légale (notamment avec le droit à l'explication explicité dans le RGPD), quand les États-Unis d'une part affichent une législation beaucoup plus souple et la Chine d'autre part développe un système de crédit social aux antipodes de la transparence, sans aucun droit de contestation. **L'interprétabilité apparaît donc protéiforme et se diversifie selon les craintes, les ambitions et la culture de chaque individu ou pays.**

GRÉGOIRE MARTINON

ILS ONT TENTÉ L'EXPÉRIENCE

CHUS

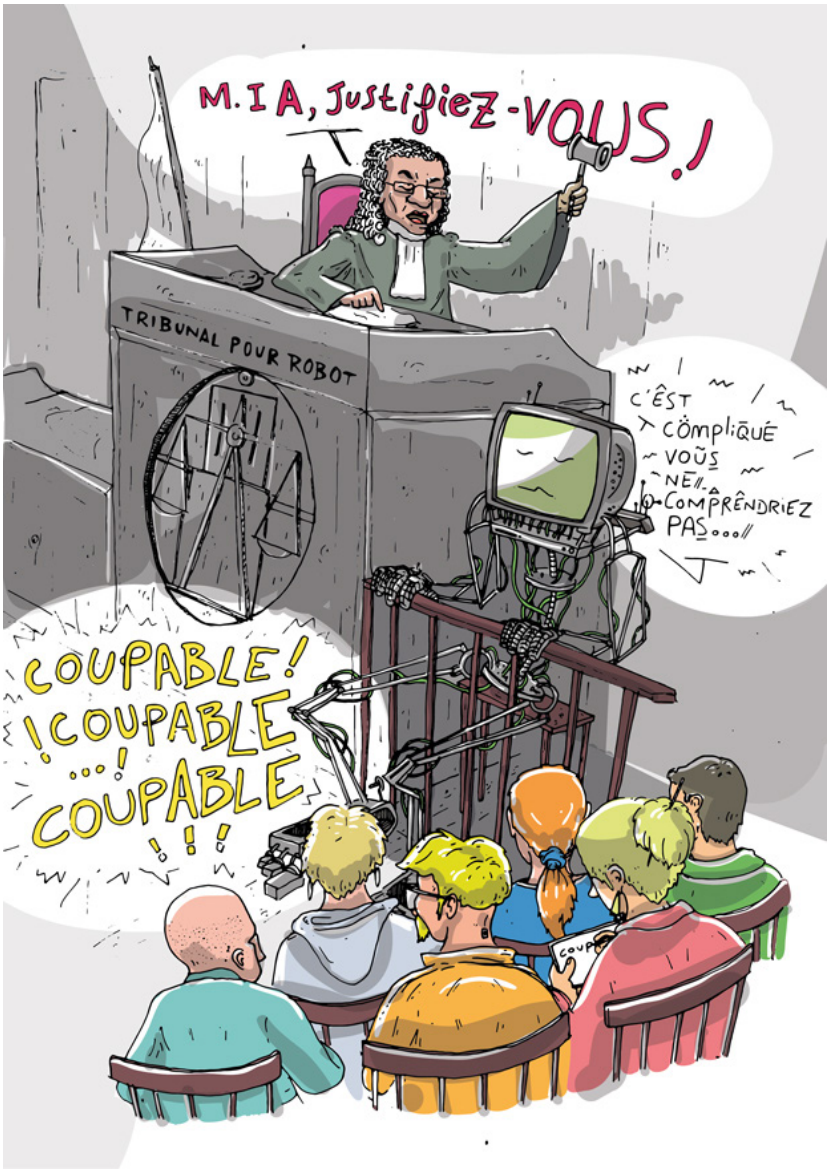
Il y a une véritable attente de la part des patients et du personnel des hôpitaux vis-à-vis de cette nouvelle médecine avec les algorithmes, mais il y aura une forte résistance de la part des patients envers l'utilisation d'un modèle si le médecin n'a pas de regard critique sur les résultats de celui-ci. L'algorithme peut faire une proposition,

mais celle-ci doit impérativement être validée par le médecin. Au quotidien l'algorithme doit être une véritable aide à la décision pour le médecin, dès lors que les critères pris en compte pour arriver au résultat sont connus de manière explicite. En effet il est possible que le médecin n'ait pas pris en compte un paramètre que l'algorithme aura lui intégré. Ces paramètres doivent être complètement transparents pour le médecin. De manière générale le fonctionnement mathématique du modèle n'est pas important pour le médecin, c'est bien la pertinence empirique qui importe.

GRTgaz

L'intérêt de l'interprétabilité est de pouvoir expliquer les résultats à la fois aux exploitants qui reçoivent les prédictions de stock et au management qui prend les décisions d'investissement sur ce projet. Il s'agit d'être capable d'expliquer pourquoi on arrive à ce résultat, sans nécessairement comprendre l'intégralité du modèle mathématique. Il faut pouvoir comparer les prédictions du modèle avec la méthode précédente plus simple afin de challenger la cohérence du modèle, et pour ce faire il est nécessaire de comprendre les variables utilisées et les règles métiers prises en compte. Il y a eu une conduite du changement progressive envers les dépôts pour les informer sur l'utilisation d'un nouveau modèle de *Machine Learning* et surtout sur l'objectivité des calculs qui sont faits par la machine. Les responsables des dépôts ont confiance dans le modèle, ils vont surtout vérifier le montant de stock préconisé, et certains vont demander les critères utilisés.

C. DE LA CONTRAINTE À L'OPPORTUNITÉ



UN ENVIRONNEMENT QUI SEMBLE PEU FAVORABLE À L'INVESTISSEMENT

Devant l'avènement des nouvelles contraintes réglementaires, la nécessité d'interprétation des modèles ne fait pas de doute. **Des craintes émergent ainsi quant à l'impact sur l'innovation en Europe dans le domaine de l'IA :**

› Les **systèmes interprétables pourraient être moins performants** et conduire à des décisions peu appropriées pour le consommateur ;

› L'IA est supposée apporter des solutions plus objectives, détecter et corriger les biais dans le jugement humain, **pas être construite sur les mêmes bases humaines qui font défaut aujourd'hui ;**

› **Les coûts liés à l'utilisation de l'IA vont augmenter** à cause des tâches supplémentaires à prévoir lors d'un nouvel usage des données : analyse des impacts, revue et contrôle manuel des décisions, adoption de technologies et d'organisations appropriées, demande de consentement ;

› **Le risque opérationnel va s'étendre** : de fortes amendes seront imposées en cas de non-respect du RGPD, même si ce dernier reste pour l'instant un texte particulièrement complexe et sujet à plusieurs interprétations possibles ;

› **Le marché européen pourrait perdre son attractivité,** les régulations étant plus clémentes dans le reste du monde.

Si ces craintes sont justifiées, il reste à voir **comment la mise en application des réglementations se réalise, car on observe pour l'instant un manque de jurisprudences** ; par rapport au degré de transparence demandé pour l'IA, en particulier, les règles édictées restent assez vagues.

DÉMYTHIFIER L'IA

Il existe beaucoup d'idées fausses vis à vis de ce que l'IA est capable de réaliser. Une clarification s'impose afin de mieux cerner les contours de l'obligation de transparence et de ce qu'elle implique.

Considérons d'abord le cas d'une IA forte, c'est à dire capable d'une intelligence humaine, d'initiatives voir de conscience. La prise de décision pourrait faire abstraction de la composante rationnelle au privilège des émotions, ce qui rendrait la notion d'i**nterprétabilité extrêmement complexe** : en effet, les raisons qui poussent un humain à prendre une décision sont elles-mêmes souvent trompeuses et impliquent notre inconscient. Le développement d'une telle IA, qui po-

serait des forts **problèmes d'acceptation par la population**, est pour le moment utopique (ou dystopique, selon les points de vue !).

Cela est à contraster avec l'IA spécialisée actuellement en usage, qui s'applique dans un cadre spécifique à la résolution d'un problème donné et qui cherche à minimiser une erreur ou, dit autrement, à maximiser la vraisemblance de ses prédictions. Les résultats, basés sur des données d'entrée connues et contrôlées, sont dans ce cas-là froidement rationnels et découlent d'une suite de calculs logiques. L'interprétabilité consiste ici à expliquer cette logique mathématique.

L'IA développée actuellement possède ainsi une nature cohérente propice à l'interprétabilité, qui laisse envisager une réduction du décalage entre la complexité des modèles mathématiques et la demande d'un raisonnement appréhendable par l'humain.

UN CATALYSEUR D'INNOVATION

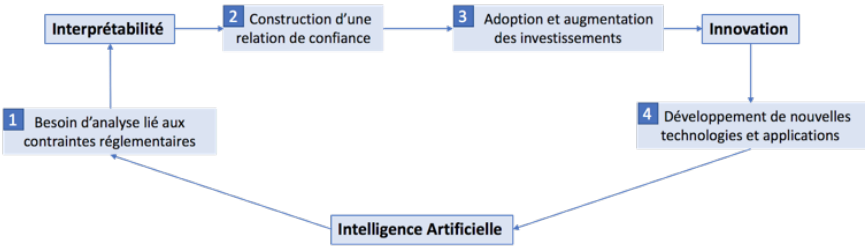


Figure 2 : Cercle vertueux de l'interprétabilité

Si les contraintes d'audit des modèles ne seront pas trop pesantes dans un premier temps, ce qui rendrait rédhibitoire l'impact des points vu précédemment sur l'écosystème lié à l'IA, **un cercle vertueux peut se mettre en place.**
L'interprétabilité pourra conduire à une démocratisation dans l'usage de l'IA, particulièrement en renforçant la confiance dans la loyauté et la pertinence des systèmes l'utilisant : la société dans son ensemble pourra davantage s'ouvrir aux opportunités apportées et une diminution du fossé technologique sera possible. Sur certains secteurs clés comme la santé ou l'agriculture, des efforts dans ce sens (comme la mise en place de labels/accréditations attestant de la fiabilité et transparence des traitements automatiques) semblent particulièrement adaptés.

Cela est essentiel pour faire accepter l'IA au sein de la société comme un outil à son service ayant comme finalités entre autres la lutte contre les maladies ou une facilitation de la circulation avec les voitures autonomes.

L'engouement de la société provoquerait une augmentation des investissements, une diversification des cas d'application et la transformation de l'IA en vecteur d'une innovation saine, orientée vers des modèles toujours plus sophistiqués mais aussi capable de toujours répondre aux exigences sur les droits fondamentaux (même si cette notion peut varier en fonction du pays concerné). En cas de réussite, des bases solides seraient posées pour faire rayonner un usage éthique et responsable de l'IA de par le monde.

Il ne faut pas oublier, en conclusion, que la recherche d'intelligibilité demande des investissements financiers et humains dans la compréhension d'un modèle outre que dans son élaboration. La commission européenne a notamment réaffirmé sa volonté d'augmenter les investissements en IA sans oublier de soutenir la recherche dans la compréhension des modèles complexes. Pensons par exemple à Alpha Go, IA produite par Deep Mind capable de battre les joueurs de GO les plus forts au monde : même si comprendre les décisions prises par cette dernière ne faisait pas partie du cahier des charges initial, Deep Mind consacre désormais une partie de son activité de recherche au questionnement de ces systèmes complexes de prise de décision.

AUGUSTIN HOFF

ILS ONT TENTÉ L'EXPÉRIENCE

BPCE

Lors d'un récent projet, il a été nécessaire de devoir simplifier les résultats trouvés avec des modèles plus complexes en des règles logiques intelligibles. En effet, la valeur ajoutée de la modélisation reposait à la fois sur le fait de pouvoir faire une prédiction automatiquement et le fait de comprendre les différents mécanismes à l'œuvre dans les modèles. Cette exigence d'obtenir des règles

logiques, d'abord perçue comme une contrainte, a été transformée en opportunité de mise en production rapide de ces règles. En effet, il n'a pas été nécessaire d'attendre qu'une infrastructure capable d'héberger le modèle complexe soit mise en service. Mais les règles logiques créées ont pu être rapidement validées et testées en production avec l'infrastructure actuelle.
A ce titre, BPCE a développé une méthodologie, skope-rules, qui permet de prédire des événements particulièrement rares et de les exprimer sous la forme d'une liste de règles logiques intelligibles. Ces travaux ont été rendus disponibles à la communauté et publiés en open source sur la librairie python scikit-learn-contrib (https://github.com/scikit-learn-contrib/skope-rules).

PSA

La majorité des projets hors véhicules autonomes chez PSA qui utilisent le *Machine Learning* sont des projets d'efficience : optimisation des processus internes, performance des campagnes marketing, contrôle qualité plus efficace avec le même nombre d'opérateurs...
Le cycle projet se compose de 4 étapes : cadrage, POC, pilote et industrialisation. On évalue le modèle à l'aide de métriques de performance scientifiques. La transformation des processus se fait après l'expérimentation, lors des phases de pilote et d'industrialisation. C'est à ce moment particulièrement que l'on a besoin d'interprétabilité. L'interprétabilité est essentielle dans le secteur automobile où il faut améliorer la prise de décision à tous les niveaux : achats, commerce, usine, ingénierie... En résumé, l'interprétabilité est la clé du P&L, c'est ce qui donne de la plus-value à travers la transformation des processus métiers.

GRTgaz

Le fait de comprendre les résultats en tant qu'utilisateur métier permet d'être en capacité de mieux les interpréter. Le modèle constitue un outil d'aide à la décision, qui requiert ensuite une validation humaine permettant de détecter d'éventuelles incohérences. En effet l'historique de données qui est utilisé par le modèle ne permet pas toujours de réaliser une prédiction correcte. Par exemple pour le modèle de prédiction des niveaux de stock des pièces de maintenance, il peut arriver qu'une pièce ne soit plus du tout utilisée en raison d'une nouvelle réglementation par exemple ou à contrario le plan de maintenance peut changer entraînant des modifications des niveaux de stock.

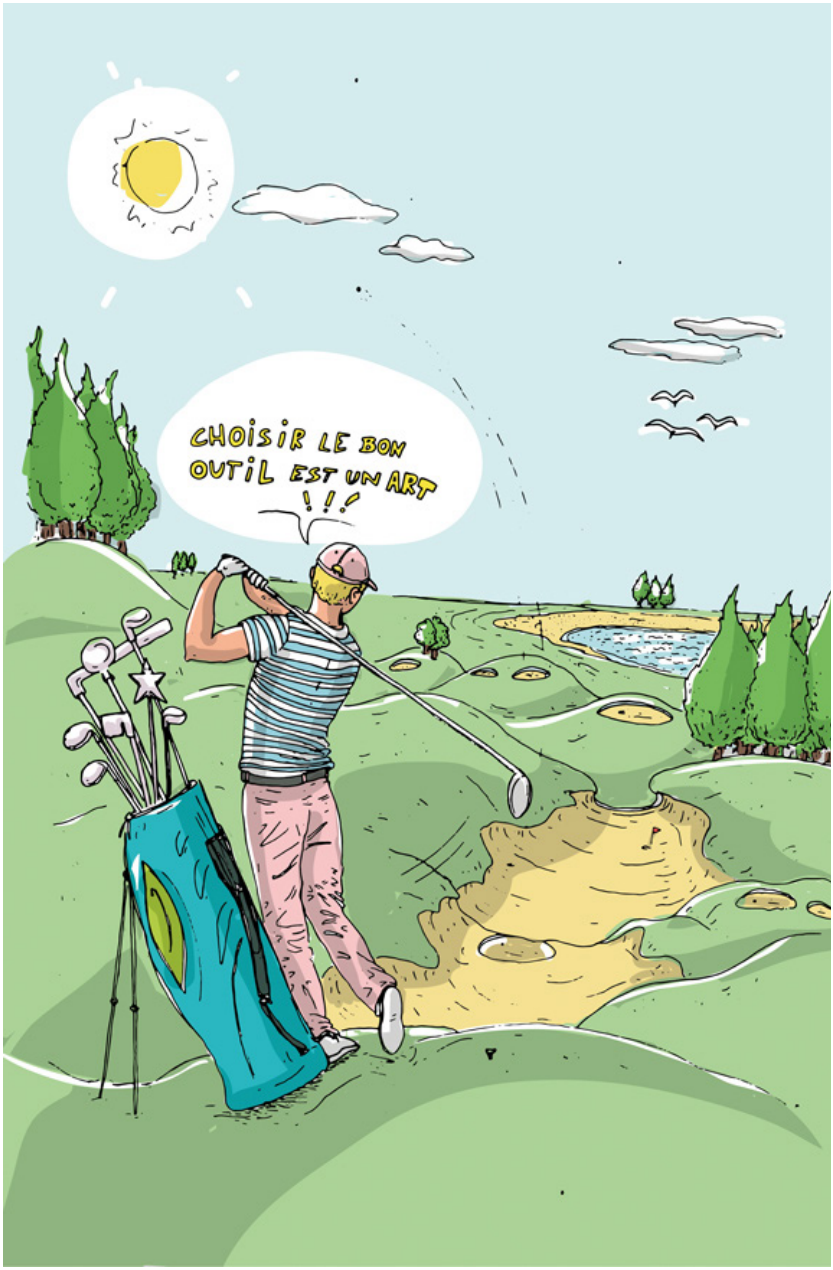
3. TECHNIQUES ET OUTILS

Il existe une multitude d'outils qui peuvent être utilisés par un *data scientist* pour mettre en place un modèle interprétable : leur pertinence dépend de plusieurs aspects tels que la nature d'interprétabilité recherchée et le type de données à disposition.

Nous nous intéresserons d'abord à l'idée selon laquelle un type de modèle est associé à une qualité d'interprétabilité, assertion à nuancer pour plusieurs raisons. Nous mettons aussi en avant la différence entre l'interprétabilité globale (du modèle en lui-même) et celle locale (d'un résultat spécifique).

Par la suite, l'utilisation de surcouches qui permettent de gagner en interprétabilité indépendamment du modèle utilisé lors d'un traitement de données tabulaires est exposée à travers d'un exemple concret. Nous montrerons aussi qu'un algorithme de *deep learning* n'est pas forcément une boîte noire, en discutant l'interprétabilité des analyses d'images effectuées via des réseaux de neurones convolutifs.

Enfin, nous aborderons les difficultés soulevées par l'ajout d'une surcouche d'interprétabilité en termes d'infrastructures et de compétences requises lors d'une interview avec Martin Le Loc, manager chez Quantmetry.


CHOISIR LE BON OUTIL EST UN ART !!!

A. MODÈLES, PERFORMANCES ET TYPES D'INTERPRÉTABILITÉ

Une croyance répandue dans le monde du *machine learning* est que l'interprétabilité d'un modèle est intrinsèquement (et inversement) liée à l'opacité dont il fait preuve pour parvenir à ses résultats. Ainsi, comme la régression linéaire est le modèle le plus basique et celui que tous les apprentis *data scientists* ont utilisé en premier, il existe une conviction partagée que ce modèle est l'un des plus interprétables (sinon le plus interprétable). De plus, il paraîtrait quasiment suspect que l'on puisse augmenter indéfiniment les performances d'un modèle (en faisant du *stacking*[11] , en ajoutant une couche supplémentaire à un réseau de neurones., etc...) sans rien sacrifier. Le data scientist apprend à toujours faire des compromis (biais versus variance, rapidité d'implémentation sur un notebook versus scalabilité du code et organisation en classes) et le compromis entre l'interprétabilité d'un modèle et sa « complexité de fonctionnement » semblerait logique à première vue.

Notre conviction est qu'il ne l'est pas tant que cela. En première approximation certes, on peut considérer qu'il y a équivalence entre un modèle et un niveau d'interprétabilité. Mais cette vision binaire pour être dépassée en montrant que l'interprétabilité du modèle dépend avant tout de notre capacité à l'associer à un ensemble d'outils de restitution adaptés. Ces derniers dépendent aussi bien du modèle et de la typologie de données que du but visé (amélioration globale du modèle, aide à la prise de décision).

PERFORMANCE VS INTERPRÉTABILITÉ, UN DILEMME À DÉPASSER

Au-delà de l'impression de logique quasi évidente, le compromis entre interprétabilité et performance est clairement évoqué dans l'ISL[12] , l'une des œuvres de référence pour tout *data scientist*. Une relation linéaire entre interprétabilité et flexibilité est aussi proposée, cette dernière notion étant associée à la capacité du modèle à capter des phénomènes complexes (interactions entre variables, non-linéarités, ...).

[11] Le stacking, ou empilement, est un moyen de combiner plusieurs modèles (généralement de types différents) : le mécanisme de combinaison est que la sortie d'un classificateur sera utilisée comme données d'apprentissage pour un autre classificateur afin d'approximer la même fonction cible.

[12] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112). New York: springer.

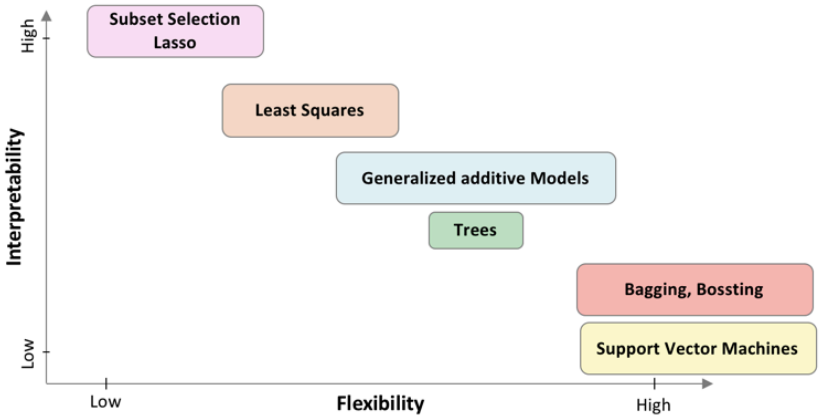


Figure 3 : Diagramme flexibility/interpretability tel que présenté dans l'ISL[12]

Cette vision bi-variée peut être dépassée en considérant davantage de facteurs qui influent sur l'interprétabilité d'un modèle :

› **Le modèle en lui-même.** Pour la régression linéaire, par exemple, on a accès directement aux paramètres qui déterminent entièrement le modèle et le rendent intelligible. Les résultats d'un réseau de neurones, au contraire, ne sont pas directement compréhensible même si on a accès à tous les poids des matrices qui le constituent ;

› **Les outils à disposition pour le rendre interprétable**. Même si un modèle n'est pas compréhensible de manière immédiate, il est souvent possible d'utiliser des outils d'interprétabilité qui lui sont propres et qui le rendent intelligible. Ainsi, il est possible de rendre une forêt aléatoire « aussi interprétable » qu'une régression en moyennant les contributions de chaque paramètre sur l'ensemble des arbres qui la composent. De même, des méthodes plus générales (dites « model agnostic ») comme LIME[13] offrent la possibilité d'associer une surcouche d'exploitation à n'importe quel modèle ;

› **Les données en entrée.** Cela serait une erreur d'oublier qu'en l'absence de données cohérentes et de qualité il est impossible d'avoir un modèle qui fonctionne, et de fait un modèle interprétable. En allant plus loin, un modèle

[13] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

qui n'utilise pas directement les données brutes, mais plutôt des agrégats de ces dernières élevés à des puissances diverses, pourra certes montrer des meilleures performances brutes mais il sera certainement plus difficile à interpréter pour un utilisateur final.

Dans son programme XAI (Explainable Artificial Intelligence), la DARPA[14] présente la possibilité d'utiliser une surcouche de *machine learning* à visée explicative afin d'augmenter l'interprétabilité d'un modèle sans en dégrader la performance.

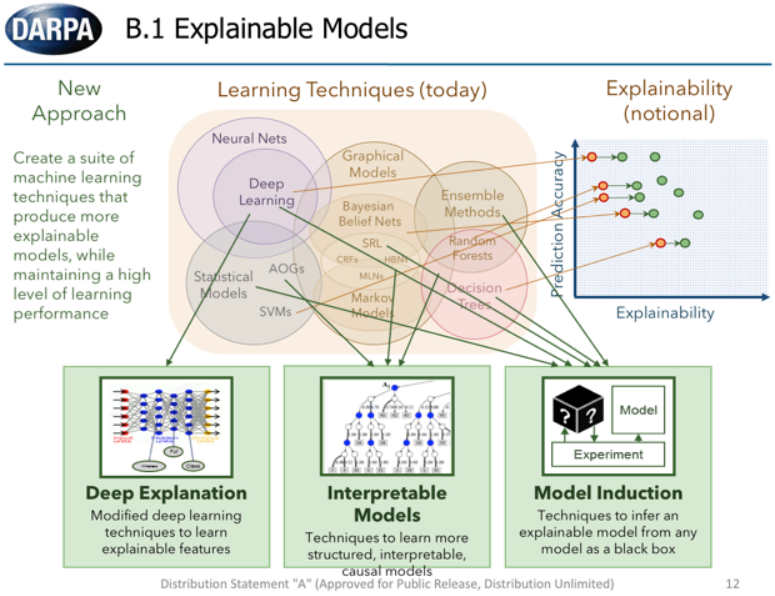
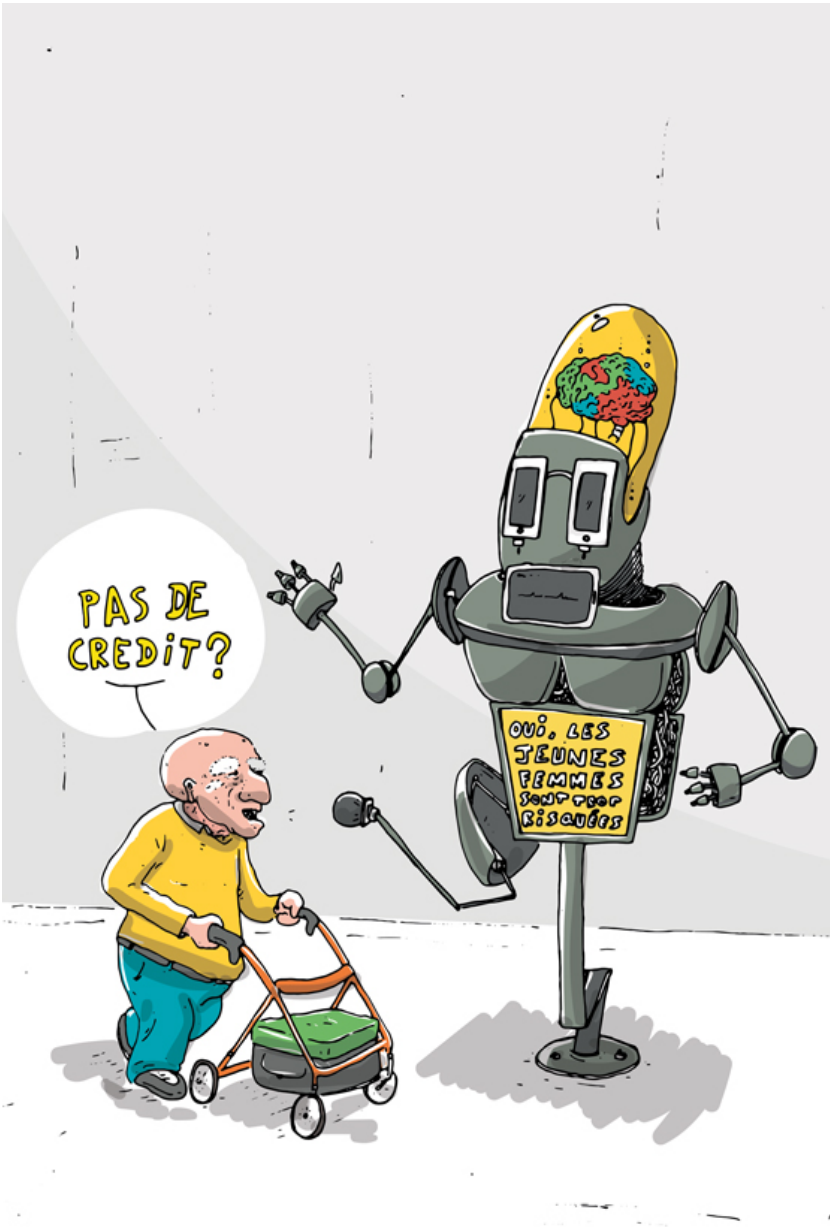


Figure 4 : Comment augmenter l'interprétabilité d'un modèle à performance constante (source : Gunning, D., Explainable Artificial Intelligence (XAI). DARPA, 2016.)

DEUX NIVEAUX D'INTELLIGIBILITÉ : LOCALE VS GLOBALE

Quand on s'intéresse à la finalité que l'on souhaite atteindre à la suite d'une meilleure compréhension du modèle, deux situations distinctes apparaissent clairement : l'interprétabilité visée peut en fait être locale ou globale.

[14] La DARPA (Defense Advanced Research Projects Agency) est une agence du département de la Défense des États-Unis responsable du développement de technologies émergentes à l'usage des militaires


PAS DE CREDIT?
OUI. LES JEUNES FEMMES sont trop RISQUÉES

L'interprétabilité globale vise à donner des indications sur comment les différentes variables d'entrée vont impacter les prédictions du modèle, alors que l'interprétabilité locale se joue à l'échelle individuelle (un individu, une machine, etc.) et vise à donner des indications sur comment les valeurs prises par les variables d'entrée impactent une prédiction isolée. Des méthodes et manières de restituer l'information différentes correspondent à ces deux niveaux d'analyse.

Pour mieux comprendre la différence, il est utile de s'intéresser à un type de modèle en particulier. Dans le cas d'une régression linéaire, le modèle est à la fois facilement interprétable localement et globalement. Considérons l'exemple hypothétique suivant :

L'efficacité d'une publicité est proportionnelle au nombre de fois qu'elle est vue par un téléspectateur et inversement proportionnelle au temps que ce dernier a passé devant la télévision.

De cette formule découle directement une interprétabilité globale : au niveau de la population dans son ensemble, la variable « temps passé devant la télévision » a un effet négatif sur l'efficacité.
La même formule se décline aussi en termes d'interprétabilité locale : afin d'augmenter l'efficacité de la publicité sur M. Dupont, le modèle suggère d'augmenter le nombre de fois qu'il regarde le spot publicitaire, ou bien de diminuer le temps qu'il passe devant la télévision.

Malheureusement, il n'est pas possible de s'appuyer sur une telle similarité entre interprétabilité locale et globale et sur la possibilité de déduire l'une de l'autre dans un cas général. On pourrait en fait imaginer un modèle plus complexe pour lequel :

- Au niveau de la population, le nombre de spots publicitaires visualisés joue un rôle important dans l'augmentation de l'efficacité de la publicité ;
- Mais pour certains téléspectateurs, qui eux auraient eu une très mauvaise expérience client avec le produit promu par la publicité, augmenter le nombre de spots publicitaires qu'ils visionnent n'a aucun effet positif sur l'efficacité de la publicité.

Il est à noter que **l'interprétabilité locale est souvent celle qui va le plus intéresser le décideur mais aussi le sujet de la décision** (qui doit pouvoir contester le fait d'avoir été classé dans une certaine catégorie par le modèle). Si l'on considère un modèle de prédiction du risque de défaut dans le cas d'un octroi de crédit, lorsque le demandeur de prêt est catégorisé comme « risqué » par le modèle, un conseiller bancaire pourra regarder quelles sont les raisons qui poussent le modèle à lui assigner ce label. Une fois cette information obtenue, il pourra la confronter à son expertise métier afin de prendre la décision la plus éclairée possible. Le cas échéant, une explication pourra être donnée au client pour le refus de son prêt. Dans ce genre de situations, une bonne interprétabilité locale est souvent suffisante.

Si en règle générale il est possible d'obtenir une interprétabilité locale de bonne qualité, il est parfois complexe d'obtenir une bonne interprétabilité globale (car celle-ci est plus corrélée à la complexité du modèle lui-même). Si le but du modèle est de trouver les déterminants d'un phénomène, cependant, elle peut être obtenue grâce par exemple à une *feature importance*[15] (dans le cas d'un modèle par arbres) ou à une analyse des neurones activés pour un input donné (pour l'analyse d'image). Les variables les plus importantes peuvent aussi être identifiées de manière itérative, en calculant à chaque étape la contribution de chacune parmi elles au score et en éliminant celle qui contribue le moins.

Ainsi, ce qui apparaît clair est que plus on complexifie un modèle, plus on augmente le travail nécessaire pour le rendre facilement compréhensible pour un utilisateur. C'est donc à ce niveau-là que se situe le compromis entre performance du modèle et interprétabilité, et le data scientist doit considérer, au-delà de la performance de son modèle, quelle sera sa possible utilisation opérationnelle. En fonction des cas, la solution la plus adaptée peut être le calcul de la *feature importance* via une brique algorithmique supplémentaire (en s'appuyant sur ses connaissances en *machine learning*) ou le développement d'une Web App d'utilisation simple grâce à une bonne UX/UI (ce qui nécessite des rudiments en théorie de présentation de l'information ainsi que des connaissances sur les librairies qui permettent de faire du front end). Dans tous les cas, il est utile de se rappeler que l'interprétabilité se traduit aussi, selon la définition donnée par Tim

[15] La *feature importance (importance* d'une caractéristique) est l'augmentation de l'erreur de prédiction du modèle après avoir permuté les valeurs de ladite caractéristique, i.e. après avoir éliminé la relation entre la caractéristique et la variable cible

Miller,[16] par la possibilité pour un humain de « prévoir les prévisions du modèle ». L'idéal est alors de donner à l'utilisateur la possibilité d'explorer le jeu de données pour vérifier ses intuitions sur les prédictions.

MARC ARTAUD DE LA FERRIÈRE

ILS ONT TENTÉ L'EXPÉRIENCE

PSA

L'interprétabilité nous montre les variables sur lesquelles on peut agir et donc les actions pertinentes à lancer. Dans le cas du scoring client, il y a certaines variables où nous n'avons pas de levier d'activation comme la distance entre le domicile du client et le garage le plus proche, et d'autres variables comme le prix pour lesquelles une action est possible. Il est essentiel d'individualiser les variables pour avoir une capacité d'action au niveau le plus fin, par client. Plus l'interprétabilité est locale, plus l'action entreprise est efficace. Un autre exemple est l'amélioration de la qualité du véhicule produit en usine : pour des défauts de peinture, l'interprétabilité a permis d'identifier l'origine du défaut qui était un problème de température.
L'interprétabilité globale ne nous dit pas comment réagir dans une situation précise. Elle a une utilité au début en expérimentation pour l'équipe de Data Science pour savoir quelles features sont pertinentes et quelles données il faut collecter.

[16] *Miller, T., 2017. Explanation in artificial intelligence: insights from the social sciences. arXiv preprint arXiv:1706.07269.*

B. LA DONNÉE STRUCTURÉE, NERF DE LA GUERRE EN ENTREPRISE

Même si des traitements de plus en plus aboutis peuvent se faire sur des données non structurées, le nerf de la guerre chez la majorité des entreprises reste l'information contenue dans leurs bases de données structurées (base clients, contrats, données de gestion, ...). **Avoir une bonne compréhension des traitements automatiques sur ces données reste donc un objectif d'importance primaire.**

UN EXEMPLE-TYPE POUR ILLUSTRER LE BESOIN ET LES TECHNIQUES

Considérons le cas d'une compagnie d'assurance généraliste *AssurGen* qui souhaite augmenter sa part de marché en assurance automobile et qui cherche à savoir comment optimiser ses leviers marketing.

AssurGen a mis en place dans le passé des campagnes de sollicitation à froid afin de proposer des contrats d'assurance automobile à ses clients. Elle aimerait en améliorer les performances à l'aide d'un score d'appétence, capable d'estimer la probabilité de chaque client de souscrire un tel contrat. Pour ce faire, les données historiques internes sur les clients d'*AssurGen* (leur âge, leur emploi, leur état civil, s'ils ont une assurance habitation ou un prêt automobile, ...) et les résultats des campagnes précédentes sont exploitées pour déterminer les clients potentiellement les plus intéressés. **Toutes ces données se présentent sous la forme d'un tableau dont les lignes représentent les individus et les colonnes les variables mentionnées précédemment.**

Les équipes opérationnelles commerciales d'*AssurGen* ont besoin d'une part d'avoir confiance dans le score d'appétence et d'autre part de savoir sur quels éléments principaux repose le score calculé afin de pouvoir adapter leur action en conséquence.

L'INTERPRÉTABILITÉ DIRECTE DE CERTAINS MODÈLES

Parfois, nous pouvons exploiter les valeurs des paramètres internes d'un modèle pour mieux en comprendre le fonctionnement (par exemple, valeur des coefficients d'une régression linéaire, valeur des coupures pour un arbre de décision...). Les modèles qui permettent cette approche sont parfois qualifiés d'interprétables.

Dans le cas d'*AssurGen*, nous pouvons utiliser une **régression logistique**, adaptée à ce type de problème, qui permet d'obtenir un score interprétable en associant à chaque variable un **coefficient** qui en quantifie l'effet sur les résultats (explication globale du modèle). Il est alors possible de construire une grille de score pour chaque client avec la **contribution** de chaque variable au score (explication locale).

Une autre méthode d'analyse reposerait sur l'**utilisation de règles basées sur un nombre de variables restreint.** Dans ce cas-là, nous serions capables de spécifier explicitement les règles de décision qui constituent le modèle. Les modèles correspondants sont alors typiquement des arbres de décisions ou des règles de décision optimisées (Bayesian Rule List, SLIM, CORELS…). Dans le cas d'*AssurGen*, une règle séparerait par exemple lors de la prédiction les assurés de moins de 35 ans, ceux de 35 à 50 ans et ceux de 50 ans et plus.

La **famille des modèles additifs** (Generalized Additive Models, Multivariate Adaptive Regression Splines (MARS), etc.) n**ous permet d'obtenir une interprétabilité semblable à celle des modèles linéaires généralisés** : chaque variable étant séparée dans l'équation de prédiction, on peut calculer la contribution de chaque variable pour une prédiction en particulier.

AJOUTER UNE SURCOUCHE D'INTERPRÉTABILITÉ À UN MODÈLE DIFFICILEMENT INTERPRÉTABLE

UN EXEMPLE DE SURCOUCHE SPÉCIFIQUE : TREEINTERPRETER

Certains modèles, considérés comme peu interprétables a priori, possèdent des surcouches propres nous permettant de mieux en comprendre le comportement.

C'est le cas pour les **forêts aléatoires**, pour lesquelles nous pouvons obtenir directement l'importance de chaque variable de manière globale et aussi, en utilisant la méthode **TreeInterpreter**[17], une bonne interprétabilité locale (avec le signe de la contribution de chaque variable pour la prédiction considérée).

[17] Palczewska, A., Palczewski, J., Robinson, R.M. and Neagu, D., 2014. Interpreting random forest classification models using a feature contribution method. In *Integration of reusable systems (*pp. 193-218). Springer, Cham.

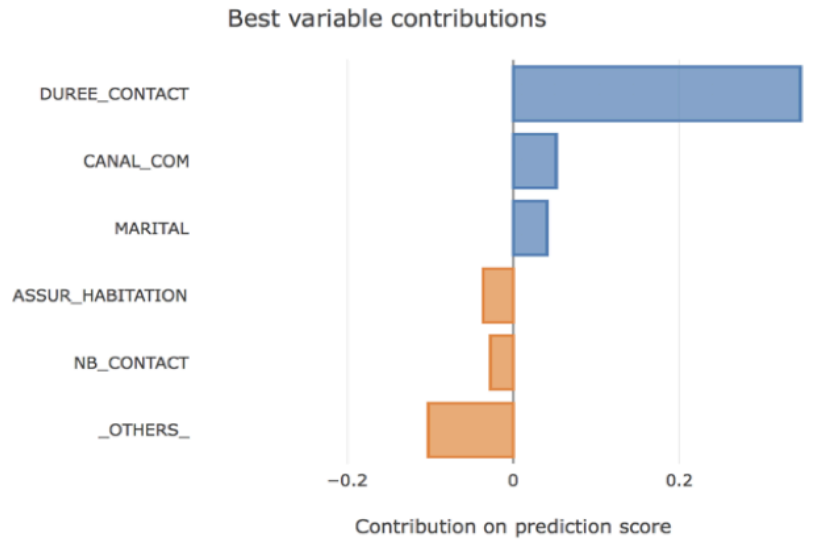


Figure 5 : Exemple d'interprétabilité locale : contributions plus significatives pour un score de prédiction donné

VISUALISATION DES PRÉDICTIONS VARIABLE PAR VARIABLE

Il est possible de visualiser l'impact de variables sur la variable à prédire une fois que le modèle a été entraîné en utilisant des partial *dependence plots.* Ceux-ci représentent la manière dont la variable à prédire évolue en fonction d'une variable de prédiction pour un modèle considéré.

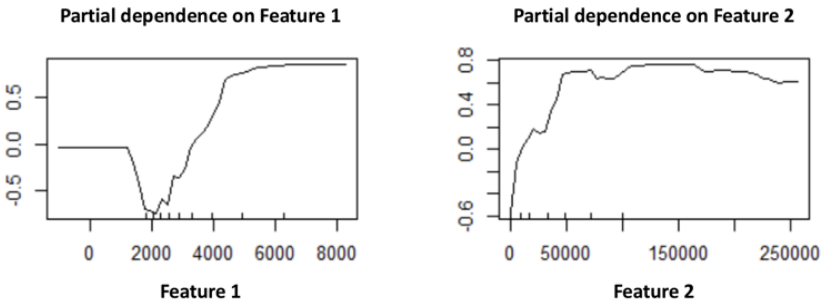


Figure 6 : Exemples de partial dependence plots pour deux variables

L'utilisation de *partial dependence plots* repose toujours sur une moyenne d'échantillons pour lesquels une variable est modifiée toutes choses égales par ailleurs. Il peut arriver que cette approche rende invisible des tendances d'évolutions particulières ayant lieu sur une sous population, auquel cas il est judicieux de préférer une visualisation sous forme d'ICE *(Individual Conditional Expectation)* curve.

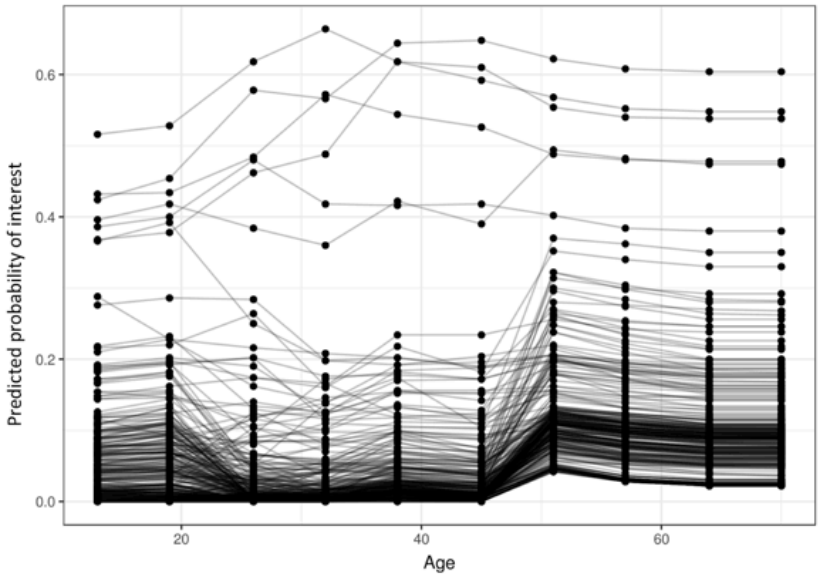


Figure 7 : Individual Conditional Expectation curve pour la variable "Âge"

Dans la même catégorie de méthodes, nous trouvons aussi l'analyse de sensitivité basée sur la variance (connue aussi sous le nom d'indice de Sobol[18]) qui décompose la variance de la sortie du modèle en fractions qui sont associées à une entrée ou à un ensemble d'entrées.

CALCUL DE CONTRIBUTIONS : LIME ET SHAP

Nous pouvons utiliser un modèle initialement peu interprétable dans le but d'obtenir des bonnes performances et lui ajouter un modèle plus explicite, qui en explique les résultats, en tant que couche additionnelle. Les effets des variables

[18] Sobol, I.M., 1993. Sensitivity estimates for nonlinear mathematical models. *Mathematical modelling and computational experiments*, 1(4), pp.407-414.

sur le score peuvent par exemple être explicités grâce à une simple régression logistique entraînée sur les résultats d'un modèle « boîte noire ».

Il faut s'assurer que le modèle explicite soit suffisamment proche du modèle boîte noire pour en tirer des conclusions. En effet, cette méthode peut avoir tendance à n'expliciter le score du premier modèle que pour un sous-ensemble restreint de l'ensemble des données. Un compromis entre l'adéquation du modèle explicite au modèle de prédiction et l'interprétabilité que l'on peut en retirer est alors à trouver.

La méthode **LIME** (*Local Interpretable Model-agnostic Explanations*) propose d'ajuster localement (i.e. sur le voisinage du point pour lequel on souhaite expliquer la prédiction) un modèle plus simple. Ainsi, pour chaque client pour lequel on souhaite expliquer le score on entraîne un modèle. Cette méthode permet d'obtenir la **contribution** de chaque variable au score sous analyse (explication locale).

En considérant l'exemple de la société *AssurGen*, il serait possible à un commercial de remarquer qu'une sous population a un intérêt très faible pour l'assurance automobile et que ce résultat s'explique très largement par l'âge relativement avancée de la population considérée. Il serait alors judicieux pour le commercial de revenir vers cette sous population en leur proposant un contrat plus adapté aux besoins des personnes âgées.

La méthode **SHAP**[19] (*SHapley Additive Explanation*) permet également d'expliquer localement le comportement d'un modèle « boîte noire ». Contrairement à LIME, où le calcul de la contribution d'une variable est défini par le modèle local interprétable choisi, SHAP fonde le calcul de la contribution d'une variable sur la valeur de Shapley : cette dernière, qui est définie en utilisant les principes de la théorie des jeux coopératifs, a comme objectif la répartition équitable de la valeur du score entre les différentes variables explicatives et en détermine leur contribution.

AMÉLIE SEGARD

[19] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *In Advances in Neural Information Processing Systems* (pp. 4765-4774

C. IMAGES ET APPRENTISSAGE PROFOND

Il y a plusieurs différences fondamentales concernant les problématiques d'intelligibilité entre données tabulaires et images : **si une *feature* (variable dans un cas, pixel dans l'autre) est une échelle adaptée pour interpréter les données tabulaires, cela n'est pas le cas pour les images, où elle représente un niveau de granularité trop fin.**
D'autres typologies de données assez complexes, comme les données sonores, les séries temporelles, ou les vidéos (suites d'image) sont elles aussi soumises à ce type de contrainte. Les problématiques de traitement de langage naturel se situent dans ce contexte dans un entre-deux, selon que les mots soient traités en tant que séquence (dans lequel cas la simple *feature* a du sens) ou plutôt via la considération de leur fréquence (agrégation de *features* nécessaire pour la compréhension).

L'approche LIME trouve rapidement ses limites quand elle est appliquée à l'analyse d'image via de l'apprentissage profond. En particulier :

- Des étapes de prétraitement sont nécessaires (agrégation en « *superpixels* »). Cette phase n'a pas qu'un coût technique : les *superpixels* restreignent en fait la suite de l'analyse d'image ;
- L'interprétation d'une image du jeu de données nécessite d'effectuer plusieurs allers-retours au sein du réseau de neurones profond ;
- Les contributions sont calculées sur un modèle local, et sont donc approximées.

Le caractère agnostique de LIME fait sa force, mais sa complexité algorithmique en rend l'utilisation en production difficile et coûteuse. S**HAP, au contraire, permet de capitaliser sur d'autres travaux propres au *deep learning*** : nous présenterons par la suite les deux extensions de SHAP les plus utilisées en analyse d'image. D'autres approches par nature associées à l'apprentissage profond, telles que les dictionnaires sémantiques dont nous parlerons, existent aussi.

DEEP SHAP (SHAP + DEEPLIFT)

DeepSHAP [20] permet d'expliquer la sortie d'un algorithme d'analyse d'image par les valeurs de Shapley des pixels. Cette méthode est une approximation de SHAP, utilisable dans le cadre de l'apprentissage profond en raison de son efficacité computationnelle accrue. En particulier, cette approche intègre certaines opérations déjà présentes dans DeepLift [21], méthode introduite afin de calculer l'importance des variables dans le cas d'un réseau de neurones.

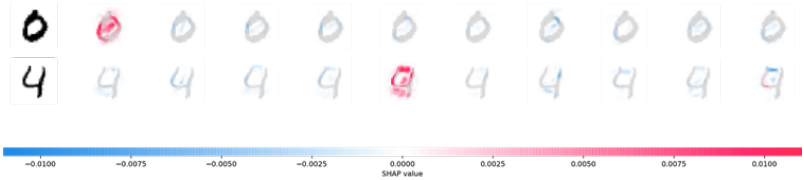


Figure 8 : Pixels importants pour les prédictions considérées (un 0 sur la première ligne, un 4 sur la deuxième)

Considérons la prédiction de classification d'un chiffre écrit à la main. En s'intéressant à la prédiction de deux images bien classées (un 0 et un 4), on peut comprendre quels pixels des images ont servis à les classifier en tant que 0 et 4 (en rouge) et quels pixels ont été considérés afin de ne pas les classifier en tant qu'autres chiffres (en bleu). Notamment, on peut remarquer que sur la deuxième ligne l'absence de barre horizontale en haut est visible en bleu dans la colonne du 9 (elle a permis de ne pas classifier l'image comme étant un 9) et en rouge dans la colonne du 4 (comme ce n'est pas un 9, le modèle comprends que c'est un 4).

EXPECTED GRADIENTS (SHAP + INTEGRATED GRADIENTS)

La méthode *Expected Gradient* étend la méthode des gradients intégrés [22] à l'ensemble des données d'entrées, et permet d'expliquer l'apport des pixels au niveau d'une couche spécifique. **Cette approche est donc à la limite entre l'expli-**

[20] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).

[21] Shrikumar, A., Greenside, P. and Kundaje, A., 2017. Learning important features through propagating activation differences. *arXiv preprint arXiv*:1704.02685.

[22] Sundararajan, M., Taly, A. and Yan, Q., 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv*:1703.01365.

cation locale (c'est bien une prédiction faite sur une image qui est expliquée) et l'explication globale (puisque c'est au niveau d'une couche entière du réseau que les gradients sont étudiés).

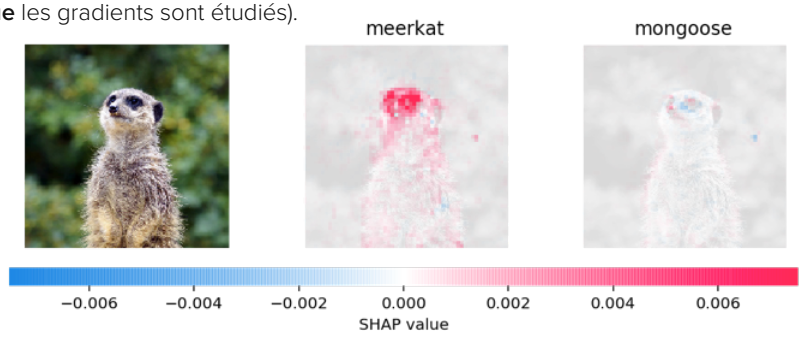


Figure 9 : Pixels importants pour le problème de classification (meerkat VS mongoose)

SEMANTIC DICTIONARIES (LUCID[23])

L'idée derrière les dictionnaires sémantiques[24] est de capitaliser sur les deux principaux apports en termes d'interprétabilité en apprentissage profond :

- La visualisation des *features,* c'est à dire la représentation graphique d'un neurone ou d'un groupe de neurones ;
- L'attribution : l'activation de cette *feature* au moment de la prédiction.

Figure 10 : Visualisation de concepts, liés à l'activation au moment de la prédiction24

[23] Python interpretability framework: https://github.com/tensorflow/lucid

[24] Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K. and Mordvintsev, A., 2018. The building blocks of interpretability. Distill, 3(3), p.e10. https://distill.pub/2018/building-blocks/

Décrire les images formées en figure 10 via des phrases serait un échec complet Décrire les images formées en figure 10 via des phrases serait un échec complet : dans l'exemple, les oreilles du chien ne ressemblent pas à des oreilles, mais plus à des formes kaléidoscopiques évoquant une oreille de chien. **On est donc face à un processus d'interprétabilité qui nécessite une couche d'interprétation supplémentaire pour en dégager un véritable sens**. Les concepts étant trop abstraits (même lorsque les neurones sont groupés), on laisse donc à l'utilisateur la « liberté » de les interpréter : cette subjectivité inhérente à la méthode n'est pas vraiment souhaitable d'un point de vue scientifique, et l'idée de décrire des images grâce à d'autres images trouve rapidement ses limites.

DES BLACK-BOX QUI COMMENCENT À S'OUVRIR

L'intelligibilité sur une image se doit d'être calquée sur la donnée d'entrée, et consiste donc dans la majorité des cas à mettre en valeur une zone de l'image pour une explication locale (tâche très bien gérée par SHAP par exemple).
Dans le même temps, il est possible de visualiser les concepts inhérents au réseau au moment de l'explication, et cela à la lumière des données passées en entrée. Contrairement à ce qu'il est possible de lire le plus souvent au sujet de l'interprétabilité de l'apprentissage profond pour les images, **il existe des méthodes permettant d'obtenir une compréhension relative de ce qui est à l'œuvre dans le réseau de neurones.** Pour autant, cette interprétabilité se résume le plus souvent à des opérations de garde-fou visant à se rassurer sur les pixels les plus importants en prédiction plutôt qu'à une compréhension véritable du fonctionnement profond du réseau.

RÉMI ADON

D. UNE MISE EN PRODUCTION À NE PAS CRAINDRE

Martin Le Loc, Manager chez Quantmetry et spécialiste d'architecture Big Data, répond à nos questions concernant les contraintes de production liées à la recherche d'une meilleure interprétabilité.

ITW : On pense souvent que l'ajout d'une surcouche d'interprétabilité à un modèle aura nécessairement un coût en termes d'organisation et d'infrastructure. Commençons par ce dernier. Quelles sont selon vous les principaux points d'attention au niveau de l'infrastructure qu'il faut avoir en tête au moment où l'on veut ajouter une surcouche d'interprétabilité à un modèle existant ou lorsque l'on veut intégrer cette surcouche à un modèle en devenir ?

Martin Le Loc : **Fondamentalement, l'infrastructure nécessaire à l'exécution d'une surcouche d'interprétabilité n'est pas différente de celle utilisée par le modèle en lui-même.** Cette mutualisation des infrastructures techniques permet de limiter le coût d'entrée d'une surcouche d'interprétabilité. Néanmoins, voici quelques bonnes pratiques à avoir en tête :

› Sur des modèles peu consommateurs de ressources : l'impact de l'interprétabilité sur l'infrastructure sera principalement sur la puissance de calcul (c'est-à-dire à la fois sur les CPUs et la RAM), le dimensionnement des ressources sera donc un point de vigilance à anticiper au plus tôt (dès le début du projet idéalement si l'on a identifié un besoin d'interprétabilité en amont) pour éviter d'impacter les performances du modèle initial notamment dans le cas de prédiction en temps réel ;

› Pour gagner en robustesse lors de l'exécution de modèles plus consommateurs, **il est également possible de séparer l'exécution du modèle d'interprétabilité de façon à isoler les deux traitements.**

ITW : Qu'entendez-vous par la séparation de ces deux traitements ? Pourriez-vous nous donner un exemple concret ?

Martin Le Loc : Imaginons l'exécution d'un modèle de prévision de défauts s'exécutant sur un environnement cloud. Le modèle d'interprétabilité ne sera pas déployé sur les mêmes ressources que le modèle initial mais sera exécuté sur un cluster de machines dédié. Il est ainsi possible d'utiliser deux fermes de serveurs

dynamiques pour moduler de façon indépendante la puissance de calcul dédiée aux deux traitements.

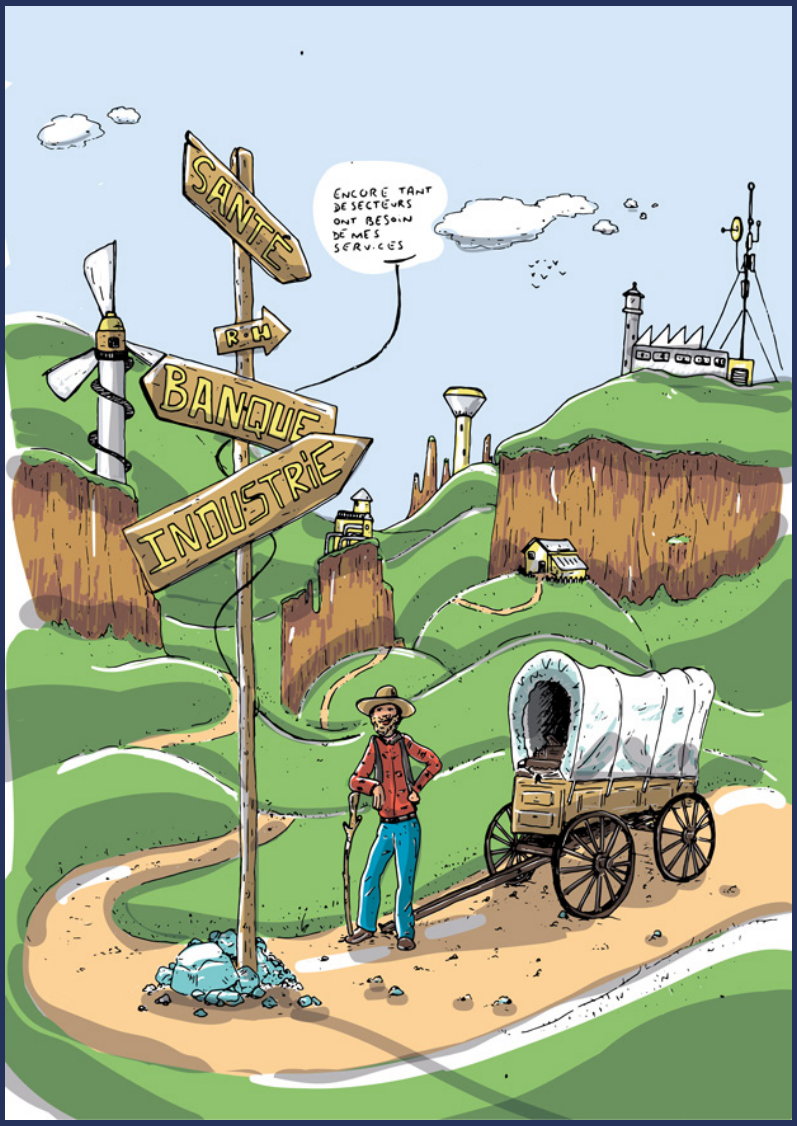
ITW : D'un point de vue technique toujours, sur quelles technologies faut-il s'appuyer pour obtenir une restitution visuelle pertinente ? Dans quel(s) cas préférez-vous mettre en place une solution sur mesure par rapport à une solution sur étagère ?

Martin Le Loc : **Nous recommandons généralement d'utiliser l'outil de visualisation de données avec lequel les utilisateurs sont les plus à l'aise (c'est-à-dire la solution retenue par le groupe et maîtrisée par une équipe**). En cas d'absence d'outil identifié ou en cas de besoin particulier (par exemple la saisie de commentaires directement dans l'outil de dashboard ou l'intégration de pièces jointes) la construction d'un tableau de bord via des outils *open source* reste une solution pertinente avec des technologies efficaces comme Dash qui permettent une livraison rapide et agile.

ITW : Côté organisationnel maintenant : diriez-vous qu'il est obligatoire d'effectuer des recrutements dédiés ou une montée en compétence de ressources existantes peut être envisageable sur des sujets d'infrastructure ou de visualisation ?

Martin Le Loc : Aujourd'hui, les expertises en interprétabilité sont rares et les besoins projet ne nécessitent pas pour autant des ressources à temps plein sur le sujet. Néanmoins, **une sensibilisation de l'ensemble des ressources existantes est primordiale, aussi bien sur les apports de l'interprétabilité que sur ses impacts opérationnels dans la réalisation des projets data.** Pour adresser ces problématiques, un référent interne et/ou une expertise ponctuelle externe sont des solutions pertinentes.

4. MA VIE DE DATA SCIENTIST : UNE APPROCHE À ADAPTER SELON LE SECTEUR



A. UNE GRILLE DE LECTURE BASÉE SUR NOTRE EXPÉRIENCE

Tout *data scientist* est confronté quotidiennement à la complexité de ses modèles et à la nécessité de les rendre interprétables pour les améliorer.

Cette nécessité, qui résulte de l'adoption de plus en plus rapide de l'IA dans les processus des entreprises ainsi que de sa régulation progressive (avec notamment la veille des *Data Protection Officers*), représente un effort d'empathie pour adapter son discours, les restitutions et les outils livrés aux différents acteurs du projet tout au long de la vie d'un projet.

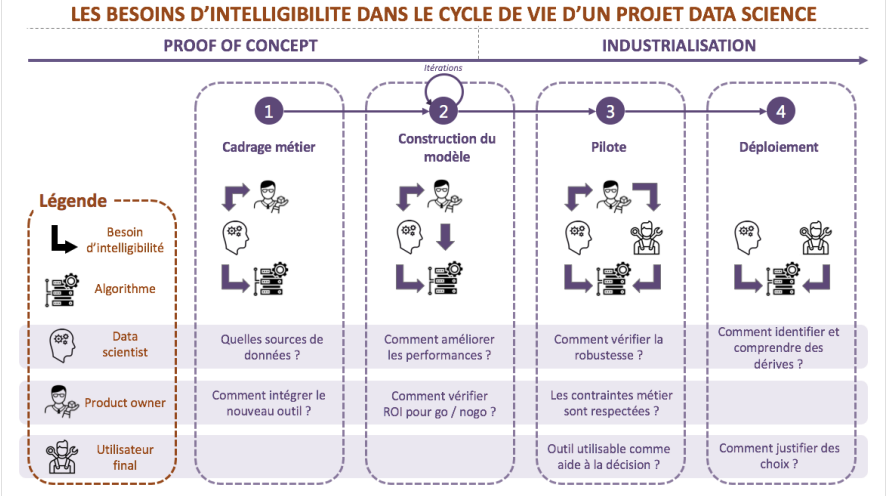


Figure 10 : Schéma explicatif des besoins en intelligibilité par les différents profils qui participent à un projet data, dans ses différentes phases

Le besoin d'intelligibilité peut être schématisé le long de deux axes *horizontal* et *vertical* :

› ***Horizontal* car il évolue au cours du temps.** De la phase initiale de cadrage avec le responsable du projet, cruciale pour comprendre les problématiques métiers et développer un climat de confiance, le discours s'adapte lors de la construction des premiers modèles sur des données de test. Comment traduire un modèle mathématique et les métriques statistiques en indicateurs utilisables par le client ? Ces modèles et métriques répondent-ils bien à son problème ? Une fois cette phase validée par un expert métier, la question de l'intelligibilité se pose à nouveau au moment du déploiement de l'algorithme

et de son utilisation potentielle par un acteur non-technique. La visualisation des sorties du modèle est-elle claire et permet-elle son adoption par tous les acteurs métiers, quel que soit leur niveau d'expertise ? Enfin, la nature même des algorithmes de *machine learning* présuppose leur fonctionnement *in fine* autonome, sans intervention du data scientist. Comment s'assurer que le modèle reste robuste, adaptable aux nouveaux flux de données sur lequel il s'entraîne et intelligible par tous une fois l'expert technique parti ?

› ***Vertical*** **car les questions d'intelligibilité se déclinent en plusieurs types pour un utilisateur donné.** Si un acteur métier doit être capable de comprendre les variables principales agissant sur la sortie de l'algorithme de manière générale (intelligibilité globale), il doit également être capable d'expliquer des résultats précis (intelligibilité locale), par exemple si le modèle doit prédire un diagnostic médical ou octroyer un crédit à un particulier. Pour le *data scientist,* ces deux niveaux de compréhension requièrent d'inclure l'intelligibilité très tôt dans le design de l'algorithme et intégrer le métier de manière continue dans les ateliers de validation pour éviter l'effet « boîte noire incompréhensible » souvent attribués aux algorithmes de *machine learning*.

L'enjeu général pour le *data scientist* est de décentrer le discours d'un point de vue purement technique et scientifique vers un point de vue compréhensible pour le métier et les utilisateurs finaux du projet.

Ce chapitre illustre cette question de l'intelligibilité à travers le retour d'expérience de quatre data scientists durant leurs missions sur des problématiques en Banque, Industrie, Santé et Ressources Humaines. Qui sont les interlocuteurs métiers ? Comment établir le dialogue et instaurer un climat de confiance tout au long du projet ? Comment adapter l'algorithme aux remarques métiers et expliquer les résultats de manière synthétique et claire pour valider son utilisation finale ? A travers ces « vies de data scientist », l'idée est loin de vouloir donner des solutions toujours valables mais plutôt de montrer comment la réflexion autour de ces sujets peut changer en fonction de l'interlocuteur, de son métier et de la spécificité du projet.

SOPHIE MONNIER

B. LE SECTEUR BANCAIRE : CASSER LES BARRIÈRES ENTRE LA TECHNIQUE ET LE MÉTIER

Au cours d'une mission dans le secteur bancaire, j'ai accompagné un client sur son premier cas d'usage *data science* et Big Data, qui consistait à mettre en place un modèle de détection d'attrition de leurs clients particuliers. Les utilisateurs finaux de ce modèle, au sein de la direction du marché des particuliers, n'avaient pas de connaissance spécifique en *machine learning* au démarrage de la mission.

Au-delà de l'aspect technique et fonctionnel de la mission, **l'interprétabilité du modèle s'est révélée être au cœur des préoccupations des différents acteurs**, et cela à plusieurs égards. Pour les experts métier, la compréhension de la construction du modèle était une question centrale : c'est pour cette raison que nous avons décidé de l**es embarquer dès le début du projet** sur tous les ateliers de modélisation, ceci permettant de discuter en continu avec l'utilisateur final et de valider avec lui de points essentiels allant de la définition de la cible (Quels clients vais-je adresser ? Quel périmètre temporel vais-je considérer ?) à la construction des variables explicatives, en prenant soin qu'elles aient un vrai sens métier, compréhensible par tous les acteurs du projet. L'inclusion du métier dans cette étape **a renforcé sa confiance dans le modèle,** apportant de fait à l'équipe data science son soutien et son adhésion au projet face aux responsables de projet.

L'interprétabilité du modèle passe aussi par le code produit par le *data scientist*. Dans cette mission, nous avons mis en œuvre une méthodologie de développement - inscrite dans les bonnes pratiques de Quantmetry – fondée sur les principes du DDD « *Domain Driven Design* [25]». L**e DDD vise à produire un code clair, orienté métier, en mettant l'accent sur la logique du domaine métier via une collaboration créative entre les experts techniques** (*les data scientists*) **et les experts du domaine** (la direction du marché des particuliers). L'idée sous-jacente est de parler le même langage entre tous les acteurs du projet, et de casser les barrières entre la technique et le métier. De fait, lors d'ateliers, l'architecture du code et ses principales fonctionnalités ont été rendues compréhensibles pour tout interlocuteur métier, ayant ou n'ayant pas d'expérience de programmation,

[25] Evans, E., 2004. Domain-driven design: tackling complexity in the heart of software. Addison-Wesley Professional.

mais également pour l'équipe *data science* de la banque, qui aura pour charge de faire vivre le modèle et de réaliser de futures itérations de modélisation. Dans ce contexte, **la clarté et la modularité du code est un atout pour l'équipe data science car elle représente à long terme un gain de productivité certain.**

Le choix du modèle résulte du compromis entre performance et complexité. Dans le cadre de cette mission, ce choix s'est porté sur les forêts aléatoires. Outre ses bonnes performances, dues principalement au travail effectué sur les variables explicatives, ce type de modèle présente l'avantage d'être facilement interprétable, à la maille globale et à la maille locale. À la maille globale, l'analyse de l'importance des variables a permis de dégager les signaux prédictifs les plus forts annonçant le départ d'un client, et de comprendre quelles sont les grandes familles de variables mobilisées pour la prise de décision.

À la maille individuelle, nous avons trouvé intéressant de construire une « user story » qui retrace l'histoire d'un client sur le départ. D'un point de vue technique, la contextualisation du modèle sur un individu représentatif de la population de clients à risque peut être réalisée avec des techniques d'interprétabilité locale telles que LIME ou SHAP permettant de décrire quelles variables ont influencé la prise de décision du modèle et avec quel poids ces variables ont affecté positivement ou négativement la prise de décision. **En utilisant ce procédé narratif, le parcours du client à risque prend sens, et permet à l'expert métier de mieux se projeter et de rapprocher ce résultat à ses observations quotidiennes**. De manière générale, l'utilisation de *user stories* pour illustrer la prise de décision d'un modèle est une bonne pratique à mettre en place quel que soit le projet.

En bout de chaîne, outre un score pour chaque client, nous avons fourni une liste de variables, issues du jeu des variables les plus importantes du modèle et complétées par d'autres variables considérées pertinentes par le métier, telles que la *durée de vie client*. **La prise de décision d'activation du modèle revient *in fine* à l'humain, qui choisira de contacter ou non un client scoré comme étant à risque.** A noter qu'un outil de visualisation des données permettant d'explorer de manière dynamique variables importantes, prévisions et explications aurait été très intéressant, car il aurait permis à l'utilisateur métier de formuler des hypothèses sur le comportement du modèle puis de les vérifier en manipulant, filtrant et explorant les résultats pour valider ou infirmer ses hypothèses.

GUILLAUME HOCHARD

ILS ONT TENTÉ L'EXPÉRIENCE

BPCE

Il est nécessaire que le métier adhère au modèle et comprenne précisément les variables d'entrées. Le modèle se développe ainsi au travers d'une relation gagnant-gagnant entre le métier et l'équipe modélisation : plus le métier s'intéresse et cherche à approfondir le modèle et à l'alimenter, plus l'équipe modélisation peut améliorer la performance du modèle ce qui le rend encore plus pertinent pour le métier. Par exemple, pour que le métier puisse valider et mettre en production un modèle de détection de fraude, il est indispensable que les stratégies de fraudes identifiées par le modèle soient compréhensibles et communiquées d'une manière intelligible.

C. LA SANTÉ : ALGORITHMES CLAIRS ET EXIGENCES STRICTES AU SERVICE DE MÉDECINS ET PATIENTS

QU'EST CE QUI REND L'INTERPRÉTABILITÉ INDISPENSABLE DANS LE DOMAINE DE LA SANTÉ ?

Le besoin d'interprétabilité des algorithmes dans le milieu de la santé vient tout d'abord du type de décisions qui en découlent. Un diagnostic, le choix d'un traitement ou la prédiction d'une rechute vont avoir un réel impact sur la qualité de vie d'un patient, voire même sur sa survie. Les processus à l'origine de décisions si critiques ne peuvent pas rester cachées à l'intérieur de boîtes noires et doivent pouvoir être appréhendées par les humains, professionnels et patients.

Par ailleurs, la notion de responsabilité de l'ensemble des acteurs de la santé publique envers la société est capitale. Qu'un algorithme maximise la rentabilité de publicités ciblées en lignes en utilisant des règles difficiles à expliciter ne pose pas les mêmes problèmes que l'attribution d'un traitement par chimiothérapie sur des critères incompris. Même s'il est entendu que tout usage de données personnelles implique un fort devoir moral, **cette responsabilité éthique devient si possible encore plus fondamentale dans le secteur de la santé**, qu'il s'agisse

d'organismes du secteur public ou privé. Nous mettons notre sort entre les mains de médecins car ils possèdent des compétences et connaissances que nous ne possédons pas ; mais cette maîtrise scientifique qui fait leur valeur et **leur légitimité ne doit pas être rongée par l'emploi d'outils au fonctionnement opaque.** Il est donc essentiel de pouvoir combiner l'information précieuse apportée par les algorithmes d'IA à l'expérience irremplaçable de l'expert humain dans l'établissement du diagnostic par le biais de processus clairs et efficaces d'interprétabilité des algorithmes.

QUEL RÔLE POUR L'INTERPRÉTABILITÉ DANS L'INTERACTION ENTRE UN MÉDECIN OU UN PATIENT AVEC L'INTELLIGENCE ARTIFICIELLE ?

Bien que les tâches que peut remplir une IA pour aider un médecin soient diverses et de plus en plus nombreuses, elles restent spécifiques et soumises à l'appréciation du médecin. Ce dernier peut ensuite considérer toutes les informations disponibles pour prendre les décisions les plus appropriées.
L'interprétabilité a donc un rôle particulièrement important dans la complémentarité entre le médecin et les outils informatiques mis à sa disposition. Sans interprétabilité, un algorithme va formuler un « avis » sans justification, difficile à exploiter par le médecin. S'il est entièrement d'accord, néanmoins, nul besoin de demander à la machine de se justifier : ce cas de figure peut être intéressant pour faire gagner un temps précieux sur certains cas d'usage, comme par exemple pour faire sortir plus rapidement le rapport d'un examen clinique qui est complètement normal. **Il serait pourtant encore plus avantageux d'utiliser les litiges potentiels pour résoudre les cas les plus complexes,** et cela n'est possible que si l'avis de l'algorithme est appuyé par une forme d'explication : l'importance d'un symptôme par rapport à un autre, la combinaison spécifique de deux indicateurs conjoints… **Le médecin pourra alors questionner son propre jugement,** le rectifier à la lueur des indications de l'ordinateur, ou au contraire comprendre l'origine d'une erreur machine.

Tirer le meilleur parti des bons résultats théoriques des algorithmes passe par l'établissement d'une confiance envers ces outils et la transparence de son processus décisionnel. La performance seule ne suffira pas à convaincre un médecin dont l'appréciation de la situation prend en compte un grand nombre de facteurs qui ne peuvent pas être passés en entrée d'algorithmes spécifiques. Le comportement du patient lors des consultations ou l'évolution de sa relation avec le médecin peuvent être utilisés par ce dernier à bon escient pour adapter le suivi.

Sans avoir confiance en l'intelligence artificielle qui l'assiste au quotidien, un médecin ne l'utilisera pas quelle que soient les améliorations qu'elle pourrait apporter.

Il en est de même pour le développement de nouvelles branches de la santé comme la médecine personnalisée. L'enjeu d'interprétabilité est d'autant plus grand que l'on se place dans un contexte où le patient se voit attribué un diagnostic directement par un algorithme d'IA, sans passer par l'avis du médecin. Outre la précision du résultat, qui est un prérequis indispensable à l'utilisation d'un tel outil, l'algorithme doit être capable **d'expliquer son diagnostic de manière exhaustive et précise, sans pour autant embrouiller le patient qui ne dispose pas du même bagage technique qu'un professionnel de santé.** La question de comment allier précision et clarté afin d'obtenir la confiance du patient est un défi de taille pour les algorithmes de demain dans ce domaine.

L'INTELLIGENCE ARTIFICIELLE COMME AIDE À LA COMPRÉHENSION DE NOUVELLES PATHOLOGIES ?

Nous souhaitons que les algorithmes puissent donner des explications en médecine. Mais si comprendre la démarche d'une machine est parfois compliqué, partager avec elle nos connaissances n'est pas toujours si simple… **En effet, de nombreux sujets en santé restent encore difficiles à expliquer ou décrire de manière détaillée :** les causes de certaines pathologies, les effets à moyen ou long terme de certains traitements, les facteurs aggravants de certaines maladies, les relations entre des grandeurs chimiques ou physiologiques mesurées, les avis contraires de différents spécialistes …

Et si l'explicabilité des modèles permettait justement de mieux comprendre ce que l'on observe ? **Plusieurs limites à la compréhension de ces exemples par les humains pourraient être levés grâce à l'usage des bons algorithmes sur les données pertinentes.** Il est notamment difficile de comprendre les phénomènes sous-jacents à des évènements rares, ou faisant intervenir un grand nombre de paramètres explicatifs, ou encore faire des estimations ou des prévisions parmi un grand nombre d'issues possibles. Au contraire un algorithme, tant que la donnée qu'il utilise est bien renseignée, peut appréhender et mettre en relation des événements éloignés dans le temps, prendre en compte un grand nombre de facteurs explicatifs et ne pas se laisser affecter par des biais qui peuvent nous convaincre à généraliser (parfois) trop rapidement ce qu'on a observé dans un faible nombre de cas.

ILS ONT TENTÉ L'EXPÉRIENCE

CHU

Le médecin doit toujours vérifier la cohérence des résultats du modèle de machine learning. Des faits cliniques supplémentaires pourront conduire le médecin à ne pas respecter les résultats de l'algorithme. Il a quoiqu'il arrive la responsabilité de l'interprétation médicale et de l'acte thérapeutique. L'algorithme est un complément, qui doit inciter la transdisciplinarité dans les hôpitaux : les data scientists co-construiront les modèles directement avec les médecins.

D. L'INDUSTRIE : UN SOCLE MÉTHODOLOGIQUE COMMUN POUR DES BESOINS VARIÉS

Nous avons eu l'opportunité d'accompagner deux grands groupes industriels dans la mise en place d'une solution *data science* visant à répondre à des problématiques métier. Ces cas d'usage consistaient **à développer des algorithmes de *machine learning* supervisé afin d'améliorer la performance des processus existants.** Dans le premier cas, l'algorithme construit à partir de données de capteurs télémétriques fournit des prédictions probabilistes pour anticiper les pannes des installations techniques. Dans le second, les prédictions permettent d'identifier les clients susceptibles de quitter le réseau d'après-vente du client en s'appuyant sur des données de CRM et d'historique de contact.

Ces deux sujets d'étude différaient aussi bien par leur secteur d'activité, leurs cas d'application et leur typologie de données mais se rejoignaient sur le besoin d'interprétabilité exprimé par les équipes métier. **Pour répondre à cette exigence, nous avons utilisé une méthodologie similaire dans les deux cas.**

L'IMPORTANCE DE CONSTRUIRE UNE RELATION DE CONFIANCE AVEC LE MÉTIER

Nos interlocuteurs principaux ne possédaient pas de connaissances spécifiques en *data science* mais une expertise technique liée au métier. Lors des différentes réunions de cadrage avec eux, nous avons pu comprendre les aspects métier cruciaux pour les procès concernés par les études. Ces facteurs ont orienté la phase d'analyse statistique ainsi que la création de nouvelles variables pour la modélisation.

Ce premier dialogue est fondamental pour la communication avec les équipes métiers et afin de créer un lien de confiance avec elles. Cela permet d'un côté au *data scientist* de comprendre les enjeux de la modélisation et le niveau d'interprétabilité exigé, et d'un autre côté au client d'être rassuré quant à la bonne compréhension du sujet par la personne qui code l'algorithme. La notion d'intelligibilité doit être bilatérale et cela passe par une compréhension réciproque des problématiques respectives.

BIEN CHOISIR MODÈLE ET MÉTRIQUES DE PERFORMANCE

Pour le *data scientist*, la question de l'interprétabilité du modèle se pose lors du choix de l'algorithme et des métriques de performance afin d'éviter une solution « boîte noire » qui complexifie la compréhension des résultats.

Pour le développement d'une modélisation prédictive plusieurs possibilités d'algorithmes et de métriques de performances sont envisageables. **Ainsi le défi est de trouver l'équilibre pour optimiser les différents critères : performance, difficulté d'implémentation et interprétabilité.** Dans les deux cas, nous avons choisi des algorithmes ensemblistes construits à partir d'arbres de décision afin d'être en mesure d'obtenir l'importance des variables utilisées pour la modélisation.
Pour le premier projet, l'algorithme XGBoost a été sélectionné : outre ses excellentes performances, il permet de mesurer l'importance des variables et de comprendre les grandes tendances de l'algorithme.
Pour le second projet nous avons évalué en parallèle la performance de plusieurs algorithmes afin de sélectionner celui présentant les meilleurs résultats. Cette démarche, menée grâce à l'outil de programmation génétique TPOT [26] ,

[26] Olson, R.S. and Moore, J.H., 2016, December. TPOT: A tree-based pipeline optimization tool for automating machine learning. *In Workshop on Automatic Machine Learning* (pp. 66-74).

nous a fait retenir ExtraTreesClassifier, un algorithme qui présente les mêmes avantages pour l'interprétabilité qu'XGBoost mais des meilleures performances pour notre cas particulier.

Le besoin d'interprétabilité entre en jeu lors du choix des métriques de performance aussi : certaines parmi elles, utiles au *data scientist* pour améliorer le modèle, sont difficilement compréhensibles d'un point de vue métier. D'un point de vue mathématique, les courbes de précision et de rappel permettent d'évaluer les performances et les améliorations d'un modèle à l'autre, et le F1 score (moyenne pondérée de la précision et du rappel) est très utilisé afin de trouver le meilleur compromis. Pour l'intelligibilité d'un point de vue métier, par contre, nous prenons le parti de présenter les performances d'une classification binaire par une matrice de confusion. **Les notions de faux positifs et vrais positifs sont directement interprétables pour un cas d'usage et permettent de rendre plus concrète une mesure de performance a priori plutôt abstraite.**

VÉRIFIER QUE L'EXPERTISE MÉTIER SOIT BIEN PRISE EN COMPTE

L'une des craintes majeures des équipes métier est que l'algorithme se trompe et omette certaines spécificités de l'expertise métier. **Rassurer sur la fiabilité de l'approche implique de renforcer la compréhension par nos interlocuteurs de la méthode scientifique employée.** La démarche pédagogique à adopter se structure en plusieurs niveaux.

Dans un premier temps, il est important d'apporter un éclairage sur la méthode de manipulation des données : quels sont les processus de préparation, de nettoyage et de formatage employés ? Garantir tout risque d'erreur et de biais dans l'apprentissage du modèle implique une phase d'analyse statistique rigoureuse. Ainsi il est primordial d'expliquer comment les données sont sélectionnées et traitées. Le choix de variables statistiquement représentatives de la population globale, l'exclusion des variables contenant l'information future, la gestion des valeurs manquantes et la normalisation sont autant d'étapes qu'il est nécessaire de présenter et de construire avec les experts métiers pour implanter une première brique d'interprétabilité.

La détection de pannes est un problème de classification binaire : il est important de s'assurer que les données présentent des comportements statistiques distincts en fonction de l'occurrence ou pas d'une panne. **Un premier atelier avec les experts métier nous a permis d'identifier les variables censées influencer la cible.** Une présentation de l'analyse statistique bivariée de ces variables permet de conforter ces hypothèses et de caractériser le comportement des installations techniques lors d'une panne.

La présentation du modèle avec les explications mathématiques de son comportement constitue la seconde brique pédagogique primordiale pour l'interprétabilité. Il s'agit dans un premier temps de présenter la *feature importance* du modèle : dégager les variables qui ont le plus de poids dans la prise de décision permet de comprendre les grandes tendances et de focaliser l'interprétation du modèle sur certaines variables clés. Ensuite, la présentation des performances par des métriques adaptées comme évoqué précédemment, qui permettent d'analyser la performance de l'algorithme et de se rassurer sur la fiabilité de l'approche proposée.

LA RESTITUTION DES RÉSULTATS

Enfin, pour illustrer à une maille plus fine le comportement de l'algorithme nous présentons quelques exemples caractéristiques de faux positifs et de faux négatifs en utilisant LIME. Cela permet d'échanger avec le métier sur l'importance des variables. **Lors d'une première restitution, notre interlocuteur s'étonnait de ne pas voir apparaître dans les variables les plus explicatives une variable qui présentait un fort impact physique sur la panne.** Avec LIME nous avons pu illustrer le comportement de l'algorithme sur des exemples et proposer un niveau d'interprétabilité plus fin qu'avec la *feature importance*. Si certaines variables n'apparaissent pas dans les dix les plus pertinentes, cela ne signifie pas que l'algorithme les exclut totalement dans sa prise de décision : **LIME permet d'illustrer ce point. De plus, LIME permet de présenter le comportement de l'algorithme lorsqu'il se trompe et apporte ainsi beaucoup de transparence,** ce qui rassure nos interlocuteurs.

La meilleure solution pour appuyer les explications sur la modélisation est de proposer un **tableau de bord interactif qui permet de fournir une visualisation de l'interprétation locale** de l'algorithme. Pour le cas d'usage de l'attrition nous

avons implémenté une interface développée par la R&D de Quantmetry, qui permet d'illustrer l'interprétabilité par des exemples concrets, et qui a été beaucoup appréciée par le client.

En résumé, la bonne pratique à adopter pour répondre aux besoins d'interprétabilité est d'assurer une communication régulière avec les métiers pour présenter via différents outils de visualisation les résultats des analyses statistiques et de la modélisation. Pour garantir un niveau d'interprétabilité suffisant, il ne faut pas hésiter à adapter sa modélisation pour converger vers un modèle plus simple mais plus interprétable. Et enfin, **pour garantir un alignement entre ces besoins d'intelligibilité et le niveau d'explication produit il est nécessaire d'aborder cette question lors des échanges hebdomadaires.**

ADÈLE GUILLET

E. *HR ANALYTICS* : AU SERVICE DE L'ÉVOLUTION DES COLLABORATEURS

Les Ressources Humaines (RH) peuvent-elles bénéficier des avancées de l'IA alors même que les sujets sensibles de l'humain sont au cœur de leurs préoccupations ? Depuis plusieurs années, de nombreuses entreprises proposent par exemple une nouvelle façon de recruter, centrée sur l'IA : ces acteurs font la promesse d'une sélection des candidatures non seulement automatique, mais aussi non biaisée, en adéquation avec les valeurs et la vision de leurs clients. Si le recrutement constitue le premier maillon de la « chaîne RH », les nombreux maillons restants sont aujourd'hui assez peu dynamisés par l'IA et ils constituent donc autant de domaines d'intérêt. C'est par exemple le cas de l'**évolution de carrière des collaborateurs.** Comment l'IA peut aider les RH dans leur suivi des collaborateurs ? Sous quelles contraintes ?

L'évolution de carrière est un élément clé de l'activité RH, en particulier depuis l'arrivée sur le marché du travail de la « génération Y ». Réputés moins attachés à une entreprise que leurs aînés, les *millenials* n'hésitent pas à démissionner et se lancer dans un nouveau challenge si l'expérience qu'ils vivent n'est pas pleinement satisfaisante. Garantir l'accomplissement qu'ils recherchent constitue donc un enjeu essentiel. Plusieurs dispositifs existent : **entretiens professionnels réguliers, bilans de compétences et entretiens annuels entre autres** sont autant d'événements pourvoyeurs de données (sous la forme de **comptes rendus écrits**), susceptibles d'ailleurs d'être combinés à d'autres sources de données, comme les **réseaux sociaux professionnels.** L'IA et les techniques de traitement automatique du langage (a.k.a. NLP) constituent des solutions appropriées pour la mise en valeur de ces documents. Sur un sujet sensible comme celui-ci, **qui touche à l'humain,** l'IA ne peut constituer une couche de décision automatique. C'est ce que décrit l'article 22 du RGPD : une personne a le droit de ne pas faire l'objet d'un traitement automatisé produisant des effets juridiques. Pour ces raisons, l'IA prend davantage la forme d'une aide à la décision pour les RH, qui peut se concrétiser par un baromètre RH « intelligent » de la santé du collaborateur au sein de l'entreprise, sur la base des **documents écrits** collectés via les différents entretiens et bilans.

Le bien-être d'un collaborateur est une première brique possible de ce baromètre RH. Son estimation repose sur de l'analyse de sentiments : à partir des do-

cuments recensés, un algorithme permet de calculer une « note de bien-être » au travail. Une question essentielle se pose alors : exploitée telle quelle, cette note a-t-elle une chance d'être pertinente ? Pas si sûr : pour un collaborateur donné, un membre de l'équipe RH a surtout besoin de comprendre **quelle combinaison de facteurs a permis de produire ladite note, et cela afin de mobiliser des leviers d'actions pertinents.** Fort de ce constat, il est indispensable de recourir à des méthodes transparentes, dédiées à assister l'utilisateur dans son processus décisionnel. Voici quelques pistes concrètes, adaptées à différents mécanismes de représentation du texte :

› Si le texte est représenté de manière non ordonnée (l'ordre des mots n'importe pas), une approche possible est de se baser sur le calcul de la matrice *TF-IDF* (ou *Bag of Words*) du corpus de documents, et sur des algorithmes de classification de type boosting / forêts aléatoires /... Ensuite, il s'agit d'extraire les (groupes de) mots qui contribuent de façon positive ou négative à une note. Si certaines méthodes, comme la régression logistique, sont directement interprétables, des outils comme *treeinterpreter* ou *LIME/Shap* sont tout désignés pour des procédures plus compliquées à interpréter ;

› Si le texte est représenté comme une séquence (l'ordre des mots importe, comme le veut l'intuition), une approche possible est de se baser sur des techniques plus avancées comme les réseaux de neurones. Dans ce cas-là, les réseaux à convolution pourraient être préférés à leurs homologues récurrents (LSTM, GRU), les paramètres de ces derniers étant plus complexes à interpréter. A l'instar des images, pour lesquelles les noyaux entraînés agissent comme des filtres (contour, forme, ...), les noyaux entraînés ici par des réseaux à convolution peuvent être alternativement interprétés comme des noyaux « positifs » (bien-être) ou « négatifs » (mal-être) afin de mettre en exergue les contributions positives ou négatives d'un (groupe de) mot de mots au sein de chaque document via des calculs simples (produit scalaire).

Les compétences d'un collaborateur peuvent constituer une seconde brique du baromètre RH. Quelles compétences un collaborateur a-t-il développé par le passé ? Lesquelles souhaite-il développer dans le futur ? Autant de questions pour lesquelles des réponses automatiques et pertinentes, sur la base desdits documents, peuvent se reposer sur la création d'un **référentiel de compétences.** Une combinaison judicieuse de plusieurs approches permet la création d'un tel référentiel : des plus simples comme la *TF-IDF*, jusqu'à des représentations sous

forme de graphes en exploitant des procédures comme *LexRank*[27] , qui utilise la notion de *salience* d'un mot dans un texte (i.e. à quel point il est important dans son contexte) et permet de réaliser une première synthèse de l'information disponible dans des **milliers d'offres d'emploi.** L'expérience montre qu'un aspect essentiel de l'utilisation de ces techniques est le post-traitement métier qui y est associé, i.e. l'ensemble des actions humaines déterministes à mettre en place, une fois les procédures statistiques exécutées. Il s'agit dans le cas présent de réunir des experts métiers capables d'extraire l'information finale (les compétences clés associées à un métier ainsi que leur hiérarchie) à partir du premier niveau de synthèse. **L'enjeu de pédagogie sur le fonctionnement des procédures algorithmiques utilisées est donc fort ;** de la compréhension de ces procédures découle un post-traitement efficace, construit en partenariat avec le métier, qui permet finalement d'aboutir à un référentiel de compétences vraiment exploitable.

D'autres briques peuvent naturellement enrichir ce baromètre. **Toutes ont en commun un élément essentiel : leur besoin élevé d'intelligibilité.** Comme elles sont utilisées pour caractériser des personnes, et potentiellement agir sur elles, leur compréhension doit être totale pour un collaborateur donné. Les deux exemples présentés illustrent des besoins à la fois en termes de prise de décision opérationnelle, par l'intermédiaire de l'ajout d'une couche logicielle expliquant les résultats de procédures algorithmiques d'IA ; ou encore en termes d'on-boarding des utilisateurs finaux, *conditio sine qua non* à la mise en place d'un référentiel de compétences pertinent. Autant de besoins qu'il est indispensable de prendre en compte lors de la mise en place de solutions ambitieuses d'IA pour le département RH.

MAYA AZOURI / ANTOINE ISNARDY

[27] Erkan, G. and Radev, D.R., 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22, pp.457-479.

5. UN LONG CHEMIN DEVANT NOUS, MAIS NOUS SOMMES DÉJÀ EN ROUTE

En 2016, Quantmetry dédia un livre blanc à l'industrialisation des projets de *data science*. Beaucoup de nos clients en étaient encore au stade de l'expérimentation plus ou moins avancée (POC, POV, pilote ...) mais notre conviction était que la valeur ne serait dégagée qu'en allant jusqu'au bout de la démarche, avec la mise en production des éclairages obtenus grâce à ces nouvelles méthodologies ; l'évolution du marché a confirmé notre vision, et les clients qui considèrent une éventuelle industrialisation de leurs projets data le seul déclencheur raisonnable pour le démarrage d'un projet sont de plus en plus nombreux. D'un certain point de vue, nous en sommes au même point aujourd'hui avec la compréhension de ce qu'il se passe dans les coulisses de l'IA. Nombre de clients commencent à se poser des questions au regard de ça, mais il est encore plutôt rare de percevoir un manque d'interprétabilité comme facteur clé pour choisir si lancer ou pas un projet, si le passer en production ou l'abandonner, si investir lourdement sur une application de machine learning ou pas. A notre avis, cela est destiné à changer au cours des années prochaines.

La nécessité de mieux comprendre les résultats obtenus grâce aux algorithmes d'intelligence artificielle est un sujet vaste, aux multiples facettes et intrinsèquement transverse.
Transverse dans les secteurs touchés : même si des spécificités existent et les différentes industries n'en sont pas au même stade d'avancement, au cours des prochaines années l'intelligence artificielle aura des impacts sur la vie de n'importe quelle entreprise (changements organisationnels, mises à niveau des démarches de prise de décision ....) ; aucun secteur pourra se tirer d'une réflexion profonde autour de ces thèmes.

Transverse dans les acteurs sociaux concernés : chacun d'entre nous est tout au long de sa vie parfois un patient, parfois un demandeur de prêt immobilier, parfois un étudiant à la recherche d'une place pour son cursus préféré…
Transverse d'un point de vue technique : la présence d'une barrière insurmontable entre boîtes noires et algorithmes compréhensibles est tout sauf que certaine ; l'effervescence dans ce domaine de recherche nous fait penser, au contraire, que le fossé encore visible à ce jour sera amené à se restreindre toujours plus.
Transverse aussi, enfin, par rapport aux expertises qui entreront dans la discussion autour de ces thèmes : loin d'être un sujet exclusivement technologique, il nous oblige à mieux définir ce que nous entendons avec les mots « compréhension », « intelligibilité », « aide à la décision »…

Le rapport Villani a eu le mérite de mettre au centre de la discussion politique ce que l'intelligence artificielle pourra nous apporter, et les tâches à accomplir pour profiter pleinement de ces perspectives. Il a aussi permis de prendre du recul, et focaliser l'attention sur une vision plus stratégique des enjeux. Les mots qu'il utilise en parlant de ce sujet (« *des systèmes autonomes […] en développant les capacités nécessaires pour observer, comprendre et auditer leur fonctionnement* ») confirment notre vision : une variété d'acteurs seront de la partie, chacun avec ses compétences, ses besoins, ses priorités.

Cette recherche pourrait bien paraître destinée à l'échec au vu de sa complexité. Au contraire, nous restons optimistes : comme on dit, la partie la plus difficile de chaque voyage est de faire le premier pas, et il ne faut pas oublier les avancées qu'ont déjà eu lieu dans ce domaine plus ou moins récemment. Les travaux de recherche scientifiques sur ces thèmes permettront aux experts métier comme aux simples citoyens de mieux comprendre ce qu'une IA peut ou ne peut pas faire, et de mieux spécifier leurs doutes et leurs craintes ; à son tour, les chercheurs et les *data scientists* pourront profiter de ces nouveaux inputs pour réorienter et rendre plus pertinents les outils développés, et une dynamique productive pourra ainsi rapidement être mise en route.

ALBERTO GUGGIOLA

Trop souvent, de nombreux projets data voient apparaître des situations où un algorithme avec des performances correctes a été créé, mais où sa mise en production pose question car son fonctionnement est difficilement compréhensible. A la fois côté recherche académique (avec entre autres la communauté eXplainable AI) et en entreprise, de nombreuses expérimentations sont menées dans le but de rendre moins opaque les modèles prédictifs créés par les équipes data.

Chez Quantmetry, nous sommes convaincus que cette démarche d'intelligibilité sera bientôt incontournable pour l'adoption de l'IA à grande échelle. Ainsi, nous avons créé un pôle d'expertise composé d'une vingtaine de consultants afin d'approfondir ce sujet et d'y dédier notre premier stepwise.

Les chapitres thématiques de notre publication, complémentaires entre eux, abordent non seulement les aspects techniques mais aussi les impacts organisationnels, réglementaires et sociétaux d'une démarche d'intelligibilité. Pour compléter cette vision, nous avons intégré des retours d'expérience venant de nos projets et des témoignages de nos clients dans plusieurs secteurs d'activité (santé, banque et assurance, industrie, énergie).

Bonne lecture !

GRTgaz Les Hôpitaux Universitaires de STRASBOURG PSA GROUPE GROUPE BPCE Coopérative, banquière et assureur autrement.