Business & Decision

DATA ÉTHIQUE

IA ÉTHIQUE

Les 2 visages d'un futur responsable

Édito

Qu'est-ce que l'éthique ? Y a-t-il un lien entre l'éthique et la data ?
Les changements de paradigmes que nous vivons actuellement, tant au niveau culturel, sociétal, social, économique, technologique et même environnemental, sont peut-être plus imbriqués et profonds qu'il n'y paraît.
Les progrès scientifiques et technologiques ont été, en à peine trois cents ans, époustouflants ; démultipliant le niveau de compréhension de l'Homme et son pouvoir d'action sur l'univers qui l'entoure. Bien sûr, personne ne souhaite finalement, ni ne peut d'ailleurs, renoncer réellement à ces progrès qui font désormais partie intégrante de notre culture. Mais force est de constater que leurs impacts réels sur nos vies n'ont pas forcément été mesurés dans toute leur ampleur et encore moins anticipés.
Que ce soit dans le milieu scientifique, technologique, biologique, et même financier, on voit de plus en plus émerger des attentes extrêmement fortes de la part des citoyen.nes qui revendiquent de nouvelles règles et interpellent les entreprises à propos de leur quête de sens et de leurs interrogations.
Le dénominateur commun de ces attentes et de ces revendications pourrait bien être l'éthique.
Dans nos sociétés modernes, la data, devenue le carburant du 21e siècle, pourrait bien avoir un lien beaucoup plus fort qu'on ne l'imagine avec l'éthique. C'est en tout cas ce que nous avons voulu montrer à travers ce livre blanc. Dans un monde de plus en plus technologique, qui s'alimente dans son processus de fonctionnement principalement de data, nous pensons que pour les entreprises, la toute première étape en matière d'éthique est d'adopter une démarche « data éthique ».

La data éthique, c'est à la fois une façon d'envisager l'utilisation des données, un état d'esprit, des valeurs. Nous pensons que la règlementation à elle seule ne suffit pas à pratiquer une data éthique. Sans autorégulation, sans convictions ni valeurs personnelles, point d'éthique !
Nous développerons des arguments qui tentent d'expliquer pourquoi les entreprises doivent s'en emparer et se l'approprier, et en matière de données, pourquoi vous devez prendre la parole et même démontrer votre éthique.
Mais pour porter ce message, nous devons revenir à la source, et nous interroger en amont et sans concession sur la question de nos valeurs.
Nous espérons que ce livre blanc vous poussera à vous engager publiquement en faveur d'une data éthique, et à communiquer sur ce sujet en interne comme en externe.
Adopter une charte accessible à tous dans l'entreprise, être transparent, assumer le positionnement. Voilà comment vous pourrez faire la différence et relier le meilleur des deux mondes : celui des technologies et celui de l'éthique. Car nous pensons qu'il est possible de créer de la valeur dans un monde résolument technologique en respectant l'Humain et l'éthique, en le mettant au centre de nos préoccupations, à condition de mettre en œuvre des actions volontaristes et d'en apporter la preuve !

Valérie Lafdal, Directrice Générale Business & Decision France
Directrice Générale déléguée Groupe Business & Decision

1

La société civile à la reconquête de la data

Pourquoi les citoyens veulent reprendre le contrôle de leurs données

Les Français affichent une méfiance croissante vis-à-vis des entreprises concernant leurs données personnelles ! 60 % des internautes se montrent en effet de plus en plus vigilants sur internet[1] (soit 6 points de plus qu'en 2017), et plus de la moitié refuse de partager certaines données et effacent leurs traces de navigation. Pourquoi une telle défiance ? Parce que l'utilisation de leurs données, notamment dans le cadre de l'intelligence artificielle, est souvent perçue comme intrusive. Preuve que les mesures et règlementations en place ne suffisent pas. Si la situation continue à empirer, le moindre dérapage – qu'il soit lié à de l'intelligence artificielle ou non – sera susceptible de créer une situation de crise, accompagnée du buzz négatif qui pourra en découler. Dans ce contexte, il devient urgent de restaurer la confiance dans les données.

Data non grata

Alors que 29 Terra-Octets de données sont publiées chaque seconde dans le monde[2], difficile de garder le contrôle sur une telle masse d'informations tant l'enjeu économique lié à leur utilisation est devenu vital pour les entreprises. Dans un contexte de compétitivité globale accrue et de course effrénée à l'innovation, les risques de dérive éthique se multiplient et s'accélèrent.

Quand la cybercriminalité devient un enjeu géopolitique

Scandale Cambridge Analytica, suspicion d'ingérence russe dans l'élection présidentielle américaine de 2016, soupçons d'espionnages technologiques de la Chine… Qui aujourd'hui a encore pleinement confiance dans l'utilisation de ses données personnelles ? À l'image du pétrole, la data – nouvel or noir du XXIe siècle et carburant de l'intelligence artificielle – représente désormais le nerf de la guerre numérique.

EN FRANCE,
92 %
DES ENTREPRISES FRANÇAISES ONT SUBI AU MOINS UNE ATTAQUE EN 2018[3].

Des crises à répétition s'inscrivent dans un contexte global de cybercriminalité accrue : avec plus de 8 millions de malwares chaque année[4], aucune entreprise n'est épargnée. Google, Uber, Facebook, Yahoo, Amazon, mais aussi British Airways, Marriott… ont toutes été victimes de fuites de données. **Depuis 2013, plus de 13 milliards de données ont ainsi été perdues ou volées dans le monde[5] !**

Un terrorisme numérique toujours plus fréquent : +32 % du nombre de cyberattaques en 2018 par rapport à 2017[6] et une tendance à la hausse en 2019 (+25 % dans les grandes entreprises rien qu'en octobre[7]). Une cybercriminalité dont le coût annuel s'élève à 600 milliards de dollars[8], soit l'équivalent de 0,8 % du PIB mondial. En conséquence : le cyberterrorisme met en exergue les lacunes actuelles quant à une protection efficace des données, renforçant ainsi le sentiment de méfiance et la frilosité des consommateurs à confier leurs informations personnelles aux entreprises.

L'avis de l'expert

Jean-Michel Franco,
directeur Marketing Produit de Talend

Il est très difficile pour un consommateur lambda aujourd'hui de faire jouer son droit à la portabilité de ses données personnelles auprès des entreprises. Nous l'avons testé chez Talend dans le cadre de notre étude Data Trust Readiness[9] et le constat est affligeant : la plupart des entreprises ne nous ont pas répondu, ou de façon incomplète et assez peu professionnelle. En revanche, dans les mentions légales, il est possible de trouver assez facilement qui contacter pour récupérer ses données. Ces demandes ne sont donc pas traitées comme un service client mais plus comme un service juridique. C'est pourquoi il reste complexe de récupérer ses données sous une forme compréhensible. Pour un consommateur, il est donc difficile de comprendre ce que font les entreprises de nos données personnelles et d'en percevoir les avantages nets.

Les équipes data n'ont pas été éduquées et manquent donc de conscience vis-à-vis de leurs responsabilités en matière de données. Pourtant le RGPD leur fournit un cadre de bonnes conduites. Au DPO alors de les informer. Voilà ce que les entreprises doivent faire en plus ! Le Privacy by Design définit des règles pour que la sécurité fasse partie de l'application et entre ainsi dans une logique vertueuse. En DataScience, il est également possible d'anonymiser les données sur lesquelles on travaille sans aucun impact sur les résultats. Ces règles doivent alors pouvoir se diffuser dans l'entreprise pour rassurer à la fois les acteurs de la donnée mais aussi le client final.

Le big data perçu comme Big Brother

La donnée est devenue un véritable enjeu de société, mais toute médaille a son revers. Celui de la data repose sur une défiance de plus en plus prononcée associée à un sentiment « d'espionnage » permanent. **57 % des internautes français se sentent ainsi davantage surveillés par les entreprises privées** (moteurs de recherche, réseaux sociaux et sites de e-commerce en tête), et **80 % jugent les algorithmes actuels des systèmes de recommandations automatiques basés sur des historiques d'achat ou de navigation trop intrusifs**[10]. En mai 2019, le sénateur démocrate américain Chris Coons exigeait lui aussi des informations quant aux enregistrements réalisés par l'assistant vocal Amazon Echo. Et que dire du système de crédit social envisagé par l'État Chinois pour « surveiller » ses citoyens en collectant leurs données ?

Quand la data met l'IA KO

Dans ce contexte, les premières applications d'intelligence artificielle n'ont fait qu'accentuer cette méfiance. À l'image d'Amazon qui a finalement dû désactiver une IA qui discriminait les candidatures de femmes à l'embauche car elle avait été entraînée avec des profils à très forte majorité masculins, ou de Tay, le chatbot de Microsoft en mode « auto-apprentissage » sur Twitter qui, en 8 heures à peine, est devenu, sous l'influence de ses utilisateurs, raciste, homophobe et négationniste. Des dérives liées à une mauvaise qualité des données, souvent incomplètes ou erronées. En ce sens, la data peut s'avérer non éthique. Une réalité qui a poussé la banque en ligne suédoise Nordnet à symboliquement « licencier » Amelia, son assistant virtuel, pour incapacité.

94%
DES INTERNAUTES SOUHAITENT REPRENDRE LE CONTRÔLE DE LEURS DONNÉES.[13]

Quelles sont alors les conséquences directes de ces dérives éthiques de la data ? À la suite de l'affaire Cambridge Analytica, **un Américain sur quatre – et même 44 % des jeunes[11] – a supprimé son application Facebook** et 54 % ont modifié leurs paramètres de sécurité. Résultat, le groupe a déjà perdu plus de 20 % de sa valeur en bourse. En Europe, le réseau social a perdu quelque 3 millions d'utilisateurs ! C'est encore peu au regard des 2,45 milliards d'utilisateurs actifs dans le monde[12], mais les comportements des clients vis-à-vis de leurs données personnelles sont désormais résolument en train de changer.

« Alors que de plus en plus de données sont aujourd'hui collectées, il est nécessaire de tendre vers un usage responsable et raisonné de la donnée à des fins business. Tout le monde doit prendre conscience de l'utilisation que l'on doit faire de la donnée pour soi mais aussi avec les autres organismes avec lesquels on interagit. »

Serge Blanc,
Data Scientist de Business & Decision

NOTES P.23

Peut-on encadrer l'usage de la data ?

74 % des Français n'ont pas confiance dans l'utilisation de leurs données personnelles par des applications mobiles[14]. Le RGPD seul ne suffit pas à changer la donne. **67 %[15] estiment en effet que le RGPD n'a pas encore permis d'augmenter le niveau de protection**. Pourquoi une telle inefficacité ? Peut-être parce que le concept même de donnée personnelle reste malgré tout encore flou, tant pour les entreprises que pour les utilisateurs.

Pourtant, il semble que RGPD offre à l'utilisateur final la possibilité de contrôler ses données, au moins en principe. En premier lieu, par **la collecte du consentement** qui doit être explicite et transparent. Deuxièmement, **par le droit d'accès et le droit de rectification**. Et, enfin, par **le droit de suppression, le droit à l'oubli** et **le droit de portabilité**, c'est-à-dire la possibilité de demander à récupérer l'ensemble de ses données personnelles à tout moment.

Mais le constat est sévère : ces obligations sont très loin d'être appliquées aujourd'hui. Un an après la mise en application du RGPD, seules 28 % des entreprises affirment être conformes et **70 % ne sont pas en mesure de transmettre les données personnelles à leurs clients lorsqu'ils en font la demande**[16], et ce malgré leur engagement formel et juridique en la matière tel que rendu public à travers leurs politiques de sécurité et de confidentialité.

« En tant qu'utilisateur, je dois d'emblée savoir que le strict minimum de mes données est collecté et, du jour au lendemain, pouvoir supprimer toutes mes données. C'est pourquoi, les utilisateurs doivent être dans la boucle de décision pour cadrer précisément ce que recouvrent les notions de data éthique et d'IA éthique. »

Romain Bernard,
Manager DataScience de Business & Decision

La data, point de départ d'une data éthique

Le RGPD est-il alors une contrainte ou un cadre minimum de ce qui devrait exister ? La data éthique est un nouveau sujet et la règlementation et la nouvelle génération font clairement apparaître des carences éthiques. Avec la montée en puissance de l'intelligence artificielle, ce sujet prend bien entendu une nouvelle ampleur. C'est pourquoi, les entreprises doivent aussi se poser la question de la finalité de la donnée. Dans la santé par exemple, la collecte et l'utilisation des données peuvent sauver des vies. La collecte, autant que la finalité du traitement, doivent donc être éthiques. La data en tant que telle peut bien évidemment être non éthique, c'est une condition nécessaire mais pas suffisante pour être éthique.

L'avis de l'expert

Stéphane Walter,
Senior Manager, Conseil, Expert Big Data
de Business & Decision

La problématique quant à la compréhension des données personnelles est réelle. Selon la CNIL, *« une donnée personnelle est toute information se rapportant à une personne physique identifiée ou identifiable. Mais, parce qu'elles concernent des personnes, celles-ci doivent en conserver la maîtrise. »*
Pour les entreprises, le fait de ne pas utiliser les données personnelles peut être un frein opérationnel pour concevoir par exemple des jeux de tests intéressants. C'est pourquoi elles peuvent être tentées de contourner cette contrainte. **Mais à moyen ou à long terme, les clients se détourneront des marques qui ne respectent pas leurs données personnelles.** C'est pourquoi les entreprises doivent entrer dans une logique éthique. En ce sens, le RGPD peut leur servir de guide, notamment face au cloud act*. Les organisations doivent prendre des mesures pour protéger leurs données vis-à-vis de tiers mais aussi de leur propre gouvernement, et assurer une utilisation éthique à leurs clients finaux. Si le client a confiance dans l'usage des données confiées à l'entreprise, il sera plus enclin à les lui communiquer.

* Le cloud act (pour *"Clarifying Lawful Overseas Use of Data Act"*) est une loi fédérale américaine promulguée le 23 mars 2018 qui contraint les opérateurs télécoms et fournisseurs de services cloud à fournir leurs informations stockées sur leurs serveurs aux « forces de l'ordre américaines ». Que ces données soient situées aux États-Unis ou à l'étranger.

78%
SONT PRÉOCCUPÉS PAR L'UTILISATION ET LA PROTECTION DE LEURS DONNÉES PERSONNELLES[17].

Tel est le paradoxe de la data : la donnée elle-même, comme les usages que l'on en fait, peuvent ne pas être éthiques. Un comportement éthique, le seul garant d'une confiance de la part des clients, implique donc non seulement une data éthique mais également les usages que l'on en fait. C'est pourquoi ni le RGPD, ni le California Consumer Privacy Act (CCPA) ne suffisent pour être éthique.
Ces règlements ont vocation à se concentrer sur la protection des données. Apple multiplie ainsi les prises de position en faveur d'une réglementation sur les données personnelle et exige désormais des développeurs d'applications de ne pas utiliser de systèmes qui enregistrent les activités des usagers sur iPhone à leur insu, sous peine de les retirer de son AppStore. Un engagement différenciant au sein de la Silicon Valley sur lequel la marque à la pomme mise fortement pour relancer les ventes de son smartphone. Même certains GAFA semblent avoir compris le potentiel marketing de l'éthique. Mais l'éthique n'est-elle pas justement le meilleur moyen de les concurrencer ?

Le chiffrement homomorphe, l'outil ultime d'un usage éthique ?

Comment alors développer des cas d'usage tout en étant éthique ? Le chiffrement homomorphe, sur lequel se positionne Orange[18], notamment, propose une approche intéressante en alliant usage et confidentialité des données manipulées.

« Ce type de chiffrement n'est pas encore répandu mais il permettra à une entité d'effectuer des calculs sur des données chiffrées, sans rien apprendre ni sur les données d'entrée, ni sur le résultat final. Mais il faut également que les entreprises respectent des codes de conduite comme celui de l'OCDE ou de l'Union européenne par exemple. Certaines entreprises envisagent d'ailleurs déjà sérieusement la mise en place d'une nouvelle fonction de « Chief Ethics Officer » sur le même principe que le DPO. Le RGPD impose le Privacy by Design, il faut faire de même avec les questions éthiques en adoptant des principes Ethics by design », insiste Mick Lévy, directeur de l'Innovation Business de Business & Decision.

2

L'éthique dans un environnement opaque

Comment faire de la data éthique dans un environnement data opaque

Quelles sont les principales causes de la défiance poussée des utilisateurs vis-à-vis de l'utilisation de leurs données ? Le constat est sans appel : un grand nombre de données sont fausses, mal collectées, non traçables, contradictoires, incomplètes et non représentatives. Les situations à la base du problème sont bien évidemment multiples, l'une d'entre elles réside par exemple dans le fait que 27 % des internautes se montrent réticents face la demande de consentement[1] et 21 % d'entre eux souhaiteraient ne partager aucune information[2]. Que font-ils alors lorsqu'ils n'ont pas d'autre choix que de renseigner leurs données personnelles ? Un tiers des Français reconnaissent fournir des informations erronées[3], voire utilisent de fausses identités ou des pseudonymes, en plus des bloqueurs de publicités[4].

État des lieux de la data : état d'urgence

Aujourd'hui, neuf entreprises sur dix estiment que leurs bases de données comportent beaucoup trop d'erreurs[5] et n'ont ainsi pas confiance dans les informations avec lesquelles elles travaillent ! Comment alors développer des applications « fiables » ou encore pire une intelligence artificielle fiable à partir de données faussées ?

« Dans un futur proche, on peut facilement imaginer que tout jeu de données présent dans un système d'information sera susceptible d'être utilisé un jour pour alimenter ou entraîner une IA », explique Didier Gaultier, directeur Data Science & AI de Business & Decision. *« Or de nombreux problèmes peuvent faire obstacle à l'entraînement correct d'une IA telles que les données fausses, les données contradictoires et les données manquantes (et/ou non représentatives). La liste est loin d'être exhaustive, le risque est alors non seulement de fausser les résultats mais aussi de reproduire voire d'aggraver les biais humains à travers des biais algorithmiques »*, insiste Didier Gaultier.

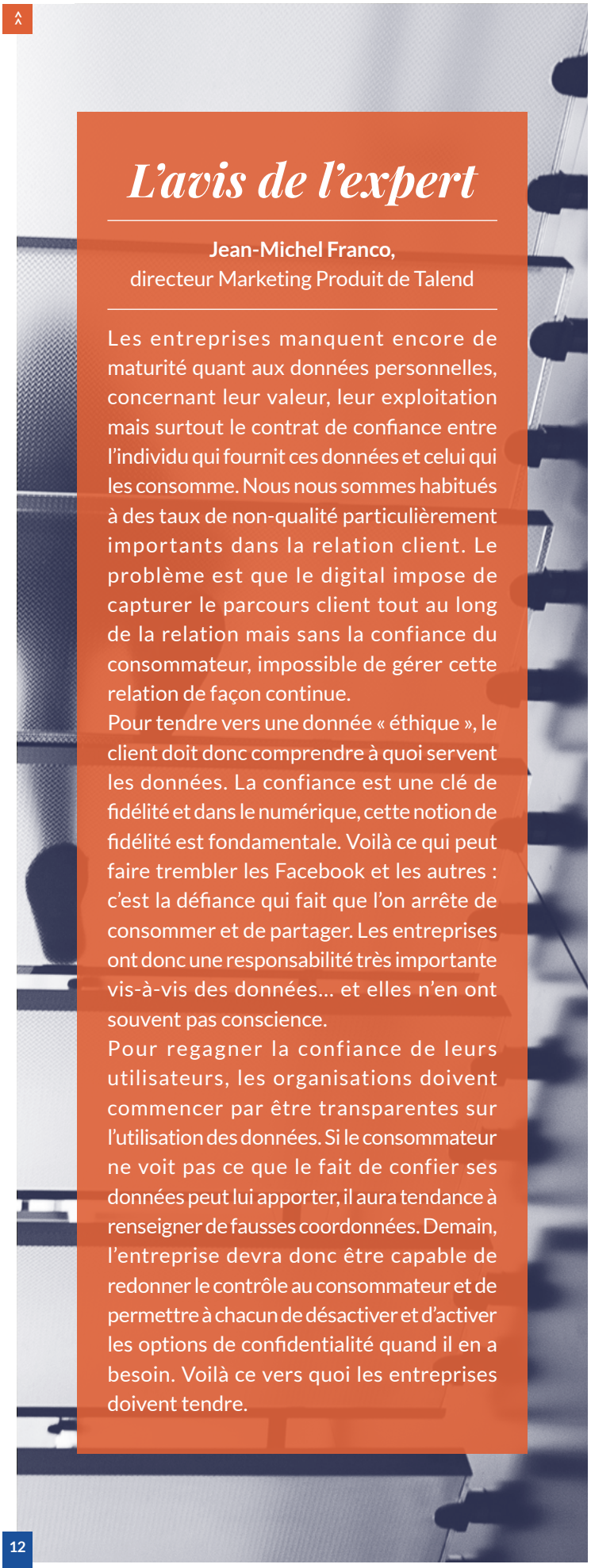
Les dix points clés d'une donnée éthique

Comment identifier si une donnée est « éthique » et digne de confiance ? Business & Decision a identifié les dix caractéristiques principales qui peuvent vous permettre de considérer n'importe quelle donnée (et jeu de données) comme éthique et fiable :

1. La data doit être définie précisément en rapport avec son ontologie[6], c'est-à-dire dans le cadre du métier que l'on souhaite modéliser.
2. Elle doit être référencée dans un dictionnaire de données accessible et à jour, incluant son nom, son type, ses caractéristiques et sa définition exacte.
3. Elle doit être juste. Dans le cas d'une incertitude, celle-ci doit alors être connue et enregistrée avec la donnée.
4. Sa date et son heure précises de collecte doivent être connues et enregistrées.
5. Son mode de collecte, intégrant les différentes sources possibles pour cette donnée (par exemple un questionnaire collecté par téléphone et par internet), doit également être renseigné.
6. La donnée doit être présente ou être explicitement déclarée comme manquante.
7. Elle doit être cohérente, c'est-à-dire qu'elle varie dans les limites définies dans le dictionnaire. De même, elle ne doit pas non plus être en contradiction avec une autre valeur liée à la même observation.
8. Elle doit être unique, à savoir qu'une observation ne doit pas donner naissance à deux entrées dans la même entité.
9. Elle doit être conforme, licite et validée, à savoir qu'elle respecte les règlements et standards de gouvernance internes ainsi que les règlements externes en vigueur (par exemple RGPD).
10. Elle doit être utile et de valeur : on ne stocke pas dans un SI des données sans avoir *a minima* un objectif d'utilisation ou de valorisation envisagé.

Très important : à partir du moment où une donnée ne satisfait pas un de ces dix points, on rentre alors dans le cadre d'une data qui s'écarte de l'éthique. Toutefois, il est bien entendu extrêmement difficile d'être à 100 % sur l'ensemble de ces caractéristiques sur 100 % des données.
Les entreprises doivent donc plutôt s'inscrire dans un « gradient éthique » qui leur permettra d'évoluer de façon continue et graduelle dans l'ensemble de ces étapes. L'idée est que ces dix points constituent un modèle, nous en convenons assez difficile à atteindre, le but étant d'évoluer vers cet idéal en ayant le plus gros pourcentage possible des données qui sont en conformité avec le plus de points possibles dans la liste ci-dessus. Le plus important n'est pas d'où vous venez, mais où vous voulez aller !

L'Apple Card, la carte de paiement du géant américain, est un exemple probant de cette complexité d'une data éthique, même pour les plus grandes marques. Tout est parti d'une série de tweets de l'entrepreneur David Heinemeier Hansson (créateur de Ruby on Rails) dénonçant le sexisme lié à cette carte, qualifiant même de *« boîte noire »* l'algorithme d'Apple. La raison ? Marié depuis de nombreuses années, il bénéficiait pourtant d'une limite de crédit vingt fois supérieure à celle de son épouse. En conséquence, les services financiers de l'État de New York ont ouvert une enquête auprès d'Apple mais aussi de Goldman Sachs, sa banque partenaire, afin de déterminer si le fonctionnement de l'Apple Card n'était pas entaché d'un biais sexiste.



L'avis de l'expert

Jean-Michel Franco,
directeur Marketing Produit de Talend

Les entreprises manquent encore de maturité quant aux données personnelles, concernant leur valeur, leur exploitation mais surtout le contrat de confiance entre l'individu qui fournit ces données et celui qui les consomme. Nous nous sommes habitués à des taux de non-qualité particulièrement importants dans la relation client. Le problème est que le digital impose de capturer le parcours client tout au long de la relation mais sans la confiance du consommateur, impossible de gérer cette relation de façon continue.
Pour tendre vers une donnée « éthique », le client doit donc comprendre à quoi servent les données. La confiance est une clé de fidélité et dans le numérique, cette notion de fidélité est fondamentale. Voilà ce qui peut faire trembler les Facebook et les autres : c'est la défiance qui fait que l'on arrête de consommer et de partager. Les entreprises ont donc une responsabilité très importante vis-à-vis des données… et elles n'en ont souvent pas conscience.
Pour regagner la confiance de leurs utilisateurs, les organisations doivent commencer par être transparentes sur l'utilisation des données. Si le consommateur ne voit pas ce que le fait de confier ses données peut lui apporter, il aura tendance à renseigner de fausses coordonnées. Demain, l'entreprise devra donc être capable de redonner le contrôle au consommateur et de permettre à chacun de désactiver et d'activer les options de confidentialité quand il en a besoin. Voilà ce vers quoi les entreprises doivent tendre.

Les problèmes de qualité les plus courants

Problème	Description du problème	Gravité	Conséquences	Remèdes
Donnée fausse	La valeur d'une observation est incorrecte ou hors des valeurs tolérées pour ce type de données (exemple client qui à 160 ans).	Potentiellement très grave suivant l'importance de la donnée surtout si elle n'est pas identifiée comme fausse.	Si elle est utilisée en entrée d'une IA, ou pire, pour son apprentissage, l'IA dysfonctionnera ou sera biaisée (Garbage In, Garbage Out).	Flagger la donnée comme fausse (ou incertaine). Corriger la valeur de la donnée si c'est encore possible.
Donnée contradictoire	Deux valeurs d'une même observation sont contradictoires dans deux enregistrements différents.	C'est le symptôme de présence de données fausses. Toutefois, elles sont alors plus facilement identifiables comme potentiellement fausses.	Si elle est utilisée en entrée d'une IA, l'IA pourrait au mieux s'arrêter de fonctionner et au pire, dysfonctionner ou être biaisée.	Flagger les deux données comme potentiellement fausses (ou incertaines). Corriger les valeurs de la donnée si c'est encore possible.
Donnée incorrectement datée dans le temps	La date de collecte (ou de traitement) d'une donnée est fausse ou manquante.	Assez grave dans la mesure où cela peut renverser un lien de cause à effet dans le processus d'inférence d'une IA ou d'un projet de DataScience.	Une IA tirera potentiellement des conclusions fausses dans un lien de cause à effet à cause par exemple d'une inversion temporelle, cela aboutira dans les cas graves à une situation de biais algorithmique.	Flagger la donnée comme incertaine ou incomplète. Compléter la valeur temporelle de la donnée si c'est encore possible.
Donnée mal définie	L'ontologie de la donnée n'a pas été définie, la donnée n'est pas référencée dans un dictionnaire de données, ou bien sa définition n'est plus à jour.	Assez grave, dans la mesure où avec une mauvaise définition, un client pourrait par exemple être à la fois éligible et non éligible à un programme de fidélité.	L'IA prendra des décisions non souhaitables. Une IA pourrait par exemple donner ou retirer à tort des droits ou des accès à un client, avec toutes les conséquences que l'on peut imaginer.	Définir une ontologie métier complète et documentée est la seule façon d'éviter ce problème. Un des résultats est un dictionnaire de données complet et à jour.
Dictionnaire de données absent	La description et la définition de tout un ensemble de données n'est pas accessible.	Suffisamment grave pour rendre les données inutilisables par les métiers par exemple.	Les données ne pourront pas être utilisées ni dans une IA ni par les Data Scientists et les DataEngineers.	Utiliser l'ontologie pour documenter un dictionnaire de données, le publier et le rendre accessible.
Mode de collecte de la donnée absent	Le mode de collecte de la donnée n'est pas renseigné (par exemple : on ne sait pas de quel type de capteur vient la donnée).	Grave uniquement s'il existe plusieurs modes de collecte possibles, par exemple par internet et par téléphone, ou bien avec différents types de capteurs.	S'il existe plusieurs modes de collectes possibles et que les données sont mélangées dans la même base, cela rend les données impropres à un traitement statistique et pourra même biaiser ou faire dysfonctionner une IA.	Isoler les data qui proviennent de sources différentes, renseigner les sources de ces données ou flagger la donnée comme incertaine si ce n'est plus possible.
Donnée manquante	La valeur d'une donnée n'est pas renseignée.	Grave seulement si c'est une donnée clé dans une analyse Data Science ou en entrée d'une IA, et suivant le pourcentage de données manquantes	Une analyse Data Science peut être rendue difficile voire impossible en fonction du nombre ou du pourcentage de données manquantes. De même, une IA peut dysfonctionner.	Compléter les données si c'est encore possible (dans certains cas, un modèle prédictif peut être utilisé), sinon flagger les données comme manquantes.
Données non représentatives	Tout un ensemble de données manquantes liées à un sujet donné rendent les données non représentatives de la réalité : par exemple, il manque toute une classe de clients.	Assez grave et insidieux surtout dans la mesure où cela n'est pas détecté.	Cela va créer un biais dans les données qui se traduira en biais algorithmique dans le fonctionnement d'une IA ou dans un projet de DataScience.	Compléter les données manquantes et s'assurer de la complétude des données. On peut pour cela procéder à des tests et des échantillonnages.
Observation dupliquée (doublon)	Deux entrées existent pour une même observation dans une même entité.	Assez grave si cela n'est pas traité.	Un client pourrait par exemple recevoir deux primes de fidélité. Un poids statistique incorrect (trop important) pourrait être attribué à un phénomène.	Supprimer la valeur en doublon par un processus de dédoublonnage.
Donnée illicite ou non conforme	La donnée n'est pas conforme à la gouvernance de données voire illicite par rapport à un règlement comme par exemple RGPD.	Grave dans la mesure où la donnée va sortir des spécifications d'entrée d'une IA et être exposée de plus à des sanctions juridiques diverses.	Conséquences potentiellement graves d'un point de vue juridique et légal. Décisions et actions non éthiques ou biaisées de l'IA.	Rendre les données conformes à la gouvernance et à la législation. Supprimer les données illicites ou non autorisées du SI.
Donnée inutile	La donnée n'est jamais utilisée dans aucune application.	Le seul vrai risque d'une donnée inutilisée est qu'elle possède un « vice caché » sans que personne ne s'en aperçoive jamais.	Prend de la place dans le SI et les sauvegardes, coûte de l'argent, des ressources et peut potentiellement avoir un problème jamais détecté.	Évaluer l'utilité réelle ou la valeur de la donnée, et la supprimer si elle n'est pas utile et n'a pas de valeur.

Voilà pourquoi une donnée peut en elle-même, intrinsèquement, être non-éthique et, dès lors, fausser non seulement les résultats des analyses opérées pas les data analystes et autres Data Scientists, mais également fausser la relation que l'entreprise entretient avec son client. La « data éthique » s'impose alors comme une condition *sine qua non* de la pérennité de l'activité.

Data éthique : une question de survie pour les entreprises

Au-delà des usages et des objectifs liés à la data, les données elles-mêmes doivent donc s'avérer éthiques et fiables au risque de biaiser les résultats et de sortir du cadre règlementaire. Les entreprises n'ont plus le choix : elles doivent donc mettre en place une gouvernance dédiée. Un écosystème certes complexe et coûteux, mais condition *sine qua non* à leur survie.

« Les coûts pour devenir éthique peuvent s'avérer très importants, confirme Cédric Missoffe, directeur de l'agence Conseil & Expertise de Business & Decision. *Chaque consentement implique un traitement spécifique. Mais si vous expliquez en toute transparence pourquoi vous collectez les emails, alors vous disposerez d'un avantage compétitif réel car la confiance accordée à votre marque n'en sera qu'amplifiée. »*

PLUS DE
95 000
PLAINTES ONT ÉTÉ DÉPOSÉES EN EUROPE POUR NON-RESPECT DU RGPD ET PRÈS DE **56 MILLIONS** D'EUROS D'AMENDES ONT DÉJÀ ÉTÉ INFLIGÉES[7].

Évolution du legacy, changement culturel profond, refonte des méthodes de travail, modification des outils de collecte et de stockage, recrutement de nouveaux profils – DPO, RSSI, Chief Data Stewards, Data Scientists, Chief Ethics Officer… Posséder de la donnée avec une finalité éthique a un coût non négligeable. Mais si vous n'investissez pas dans une donnée de qualité, le prix à payer sera bien supérieur.

L'avis de l'expert

Emmanuel Dubois,
cofondateur d'Indexima

Ces chantiers impliquent une transformation profonde, complexe, coûteuse… qui doit être portée par les directions. Or de nombreuses organisations se contentent du niveau superficiel en créant par exemple des data labs… qui souvent ne leur apportent rien. Ce n'est pas un directeur de l'Innovation qui va pouvoir à lui seul transformer une approche centrée produit établie depuis des dizaines d'années. Comment peut-on envisager de transformer une organisation silotée dans laquelle les acteurs en charge du data flow n'échangent pas et ne partagent pas la donnée entre eux ? Les freins demeurent donc principalement politiques et relèvent souvent d'une compétitivité interne entre les directions. En revanche, les organisations capables de mettre autour de la table l'ensemble des responsables de la collecte des données voient leurs projets avancer très vite grâce à des prises de décision elles aussi très rapides.
De plus, pour faire de la data un facteur différenciant, la donnée doit être de qualité et exploitable par le plus grand nombre. Plus la donnée est partagée, moins elle peut faire l'objet de biais dans son interprétation. C'est pourquoi elle doit être maturée par le plus grand nombre et circuler dans l'entreprise pour être unifiée, cohérente, traçable, croisée, utilisable et challengée par une data team qui pourra alors extraire les informations les plus pertinentes, imaginer de nouveaux circuits et améliorer ainsi l'expérience client.

NOTES P.23

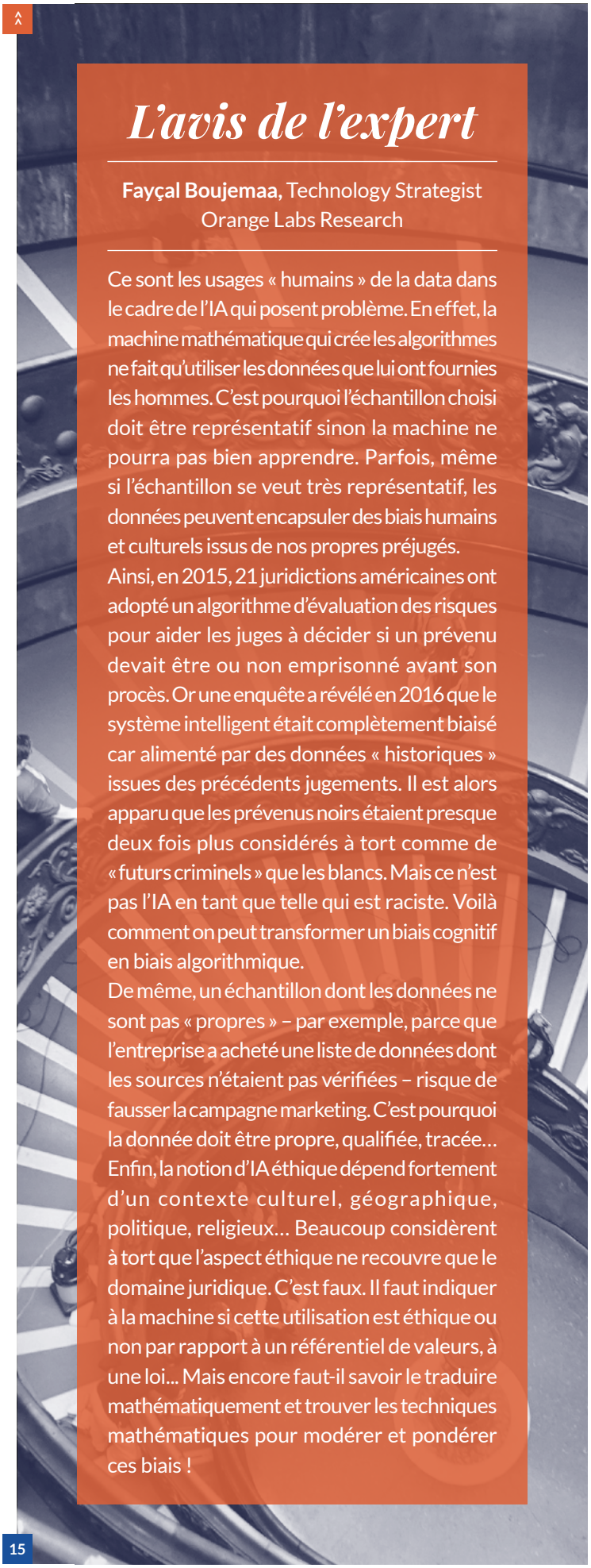
D'une approche « data centric » à une entreprise « ethic centric »

Pour tendre vers un modèle data éthique, il ne suffit pas simplement de mettre en place un simple cadre règlementaire. C'est l'entreprise tout entière qui doit tendre vers cette éthique et s'organiser autour de la donnée. Sur quel modèle repose alors l'entreprise data éthique ? Sur une équipe composée *a minima* d'un représentant métier, d'un DataEngineer pour la collecte, le stockage et le traitement des données, d'un Data Scientist pour la partie algorithmique (statistique, machine learning et IA), et d'un Product Owner comme un Chief Data Officer. Le tout sous l'œil attentif d'un « Chief Ethics Officer » capable de déployer une vision éthique globale du résultat souhaité et de faire le lien entre les différentes parties prenantes en matière d'éthique.

« *Les organismes doivent considérer leurs données comme un actif de l'entreprise au même titre que leurs ressources humaines*, explique Mick Lévy, directeur de l'Innovation business de Business & Decision. *Les données impliquent de déployer une gouvernance centralisée autour d'un Chief Data Officer par exemple, ce qui amène à poser la question de l'acculturation autour de la donnée. Il faut former, sensibiliser et accompagner les équipes internes comme les clients.* »

Pourquoi adopter une gouvernance data éthique transverse ? Parce que plus la donnée sera partagée entre des profils diversifiés, plus le risque de voir subsister des biais diminuera. Et force est de constater que les données actuelles manquent de diversité et de genre. Notamment : le secteur du numérique ne compte que 33 % de femmes dont seulement 16 % en développement et 27 % en codage[8]. Certaines écoles dont la spécialité est l'intelligence artificielle comportent actuellement jusqu'à 80 % d'hommes. Difficile dans ces conditions de certifier que les intelligences conçues sont 100 % non sexistes même avec la meilleure bonne volonté du monde !

C'est pourquoi il est nécessaire de constamment se poser la question des biais et de mettre en place des structures de gouvernance qui associent intelligence artificielle et intelligence humaine.



L'avis de l'expert

Fayçal Boujemaa, Technology Strategist
Orange Labs Research

Ce sont les usages « humains » de la data dans le cadre de l'IA qui posent problème. En effet, la machine mathématique qui crée les algorithmes ne fait qu'utiliser les données que lui ont fournies les hommes. C'est pourquoi l'échantillon choisi doit être représentatif sinon la machine ne pourra pas bien apprendre. Parfois, même si l'échantillon se veut très représentatif, les données peuvent encapsuler des biais humains et culturels issus de nos propres préjugés.
Ainsi, en 2015, 21 juridictions américaines ont adopté un algorithme d'évaluation des risques pour aider les juges à décider si un prévenu devait être ou non emprisonné avant son procès. Or une enquête a révélé en 2016 que le système intelligent était complètement biaisé car alimenté par des données « historiques » issues des précédents jugements. Il est alors apparu que les prévenus noirs étaient presque deux fois plus considérés à tort comme de « futurs criminels » que les blancs. Mais ce n'est pas l'IA en tant que telle qui est raciste. Voilà comment on peut transformer un biais cognitif en biais algorithmique.
De même, un échantillon dont les données ne sont pas « propres » – par exemple, parce que l'entreprise a acheté une liste de données dont les sources n'étaient pas vérifiées – risque de fausser la campagne marketing. C'est pourquoi la donnée doit être propre, qualifiée, tracée…
Enfin, la notion d'IA éthique dépend fortement d'un contexte culturel, géographique, politique, religieux… Beaucoup considèrent à tort que l'aspect éthique ne recouvre que le domaine juridique. C'est faux. Il faut indiquer à la machine si cette utilisation est éthique ou non par rapport à un référentiel de valeurs, à une loi... Mais encore faut-il savoir le traduire mathématiquement et trouver les techniques mathématiques pour modérer et pondérer ces biais !

NOTES P.23

Quand data éthique rime avec « customer centric »

Indissociable de l'approche data centric et d'une stratégie customer centric, la data éthique représente une transformation culturelle profonde. En ce sens, elle ne peut rester un simple élément technique et juridique, au risque d'échouer. La data éthique impose une connaissance profonde de la donnée et de l'ensemble de ses impacts.

« *La notion de customer centric était souvent perçue simplement comme un fait marketing*, analyse Serge Blanc, Data Scientist. *Aujourd'hui, on ne parle plus de produit mais de customer story et de customer experience. Donc le changement est fondamental. Tout ce qui va créer de la richesse repose sur la donnée et sur sa valeur. Cela implique de mieux comprendre son environnement, puis d'améliorer l'existant et ensuite d'innover. C'est-à-dire de créer de nouveaux services et surtout de nouveaux modèles économiques dont le point de départ repose sur la donnée. Une transformation qui impose donc de partager la donnée entre le plus de personnes possibles : voilà la richesse de la donnée.* »

Au sein d'une entreprise de plus en plus horizontale et nécessairement désilotée, il est donc essentiel de donner la possibilité à tous les acteurs d'expérimenter et d'exploiter la donnée. L'éthique commence ainsi par une prise de conscience collective qui implique de sensibiliser chaque collaborateur et d'expliquer précisément les do et les don't au sein d'une charte ou d'un référentiel de bonnes pratiques.

Faire la démonstration de l'éthique

74 % des Français n'ont pas confiance dans l'utilisation de leurs données personnelles par des applications mobiles et 78 % affirment être préoccupés par l'utilisation et la protection de leurs données personnelles[9]. Voilà pourquoi, chaque entreprise doit adopter une charte éthique interne d'utilisation de la donnée à laquelle doit se référer tout Data Scientist, un document qui engage l'entreprise et se pose alors comme un contrat moral entre l'organisme et l'utilisateur final.

Serment d'Hippocrate pour Data Scientist… ou pour toute personne travaillant avec la donnée – Data for Good

Objectif : amener le Data Scientist à prendre du recul, le responsabiliser par rapport à sa pratique et lui donner les clés pour comprendre précisément les conséquences de ses actes. Le Data Scientist doit se montrer très prudent et holistique quant à sa façon de poser les questions, et de présenter les faits de manière transparente et objective. En ce sens, l'éthique demande un certain courage et implique le devoir d'alerte.
https://hippocrate.tech/

L'avis de l'expert

Luc Julia,
CTO de Samsung Electronics, cocréateur de Siri
et auteur de « L'intelligence artificielle n'existe pas » - First Éditions - 2019

Comme tous les outils, on peut se servir de l'IA à mauvais escient. Mais ce n'est pas l'IA qui va décider d'être mauvaise, c'est nous, les humains qui allons décider de l'utiliser à mauvais escient.
Il est donc faux de dire que les outils prennent la main. En revanche, il est possible de les programmer pour qu'ils agissent mal. C'est pourquoi l'IA renvoie à l'éthique personnelle de chacun. C'est la communauté qui décide peu à peu par sa propre éthique d'une régulation, d'une loi, d'une règlementation… à l'image du traité international sur la non-prolifération des armes nucléaires en 1968 ou du Traité russo-américain sur les forces nucléaires à portée intermédiaire signé en 1987. Pour l'IA, le principe est le même : en cas de dérives, la communauté va imposer des carcans à l'utilisation de ces outils.

Mais pour que les utilisateurs comprennent, il faut une certaine éducation du public et arrêter de relayer tout et n'importe quoi au sujet de l'IA. Une fois éduqués, les consommateurs réclameront cette éthique. Je crois à la valeur de l'humain qui va prendre ses décisions en croyant ce qui est bien. C'est pourquoi être éthique représente une vraie opportunité pour une marque tout comme le concept de privacy des données dans le RGPD a eu une vraie vertu éducative, alors même que certaines entreprises l'avaient combattu au départ. Il est positif de monter que l'on est éthiquement responsable.

Pour pratiquer une « data éthique », la donnée doit alors être non biaisée. Quand on crée un système, il doit donc se montrer le plus égalitaire possible entre les différentes populations. Par exemple, si mon IA trie les CV, le data set doit avoir été choisi dans un data set équitable. Le problème de l'IA aujourd'hui est qu'elle est censée simplifier les processus, les décisions… grâce à la data, impliquant alors de disposer d'une data la plus exhaustive possible et ça, c'est compliqué. Aujourd'hui, faire des algorithmes n'est pas difficile, ce qui est complexe, c'est la data.
Donc une IA équitable pour le plus grand nombre doit provenir des data bien choisies. Voilà donc où se situe l'éthique : dans le choix de la data !
Le problème provient des données et des biais qu'elles intègrent car il y a toujours eu des bugs dans les algorithmes et il y en aura toujours. Mais les biais excluent des individus ce qui oblige alors à montrer ses data sets et à pouvoir y accéder en permanence pour corriger les bugs.

Protégez votre e(thic)-réputation

Voilà pourquoi vous ne pouvez plus faire l'économie de cette démarche data éthique au risque de vous retrouver complètement marginalisé ! *« Désormais, l'éthique dépasse le cadre de l'entreprise elle-même*, précise Valérie Lafdal, Directrice Générale Business & Decision France et Directrice Générale déléguée Groupe Business & Decision. *La passerelle actuelle entre l'écosystème des entreprises et l'écosystème social n'a jamais été aussi fine à l'image du mouvement #MeToo qui a poussé les entreprises à s'engager. Donc affirmez que vous êtes en cohérence avec les attentes de la société, revendiquez votre éthique et démontrez-la en l'incarnant au quotidien. C'est en affirmant votre identité que vous pourrez vous différencier car, demain, de plus en plus de marques disparaîtront du fait de leur mauvaise réputation. »*
Et si le prochain #MeToo était (data) éthique ?

69%
DES JEUNES CONSOMMATEURS SE MONTRENT TRÈS ATTENTIFS AUX CAMPAGNES DE COMMUNICATION D'UNE ENTREPRISE ENGAGÉE (VIS-À-VIS DES ENJEUX SOCIÉTAUX)[10] ET SE MONTRENT PLUS SUSCEPTIBLES D'ACHETER AUPRÈS D'UNE MARQUE QUI FAIT PREUVE DE TRANSPARENCE SUR SES ENGAGEMENTS.

NOTES P.23

3

L'éthique, nouveau levier de compétitivité

L'éthique, nouveau levier de compétitivité

C'est un fait, plus de sept consommateurs sur dix (73 %) prennent désormais leur décision d'achat après avoir consulté jusqu'à six avis client[1]. Avis auxquels 68 % d'entre eux font confiance, soit deux fois plus que pour une publicité issue des médias traditionnels. À l'inverse, une marque mal notée peut s'avérer disqualifiée pour 87 % d'entre eux. Tout est dit. Voilà le nerf de la guerre commerciale pour les entreprises : la confiance.

Data Wars : la guerre des data aura-t-elle lieu ?

92 % des entreprises conformes au RGPD – et donc en ce sens à une certaine approche éthique de la donnée – affirment avoir obtenu un avantage compétitif[2] alors qu'elles n'étaient que 28 % à attendre un tel avantage avant la mise en œuvre du règlement européen. Parmi les principales améliorations constatées : la confiance client (84 %), l'image de marque (81 %) et la motivation des collaborateurs (79 %). Autres facteurs bénéfiques observés de la conformité des données collectées : le renforcement de la cybersécurité (91 %) et une transformation organisationnelle (89 %) plus poussée.

Pourquoi un tel facteur de compétitivité ?
Parce que, comme l'explique **Cédric Missoffe**,
directeur de l'agence Conseil & Expertise chez Business & Decision,

« les entreprises européennes ne disposent pas des mêmes armes juridiques que leurs concurrentes américaines ou chinoises. La différence ne s'exprime pas par le prisme technologique car nous utilisons tous les mêmes intelligences artificielles. En revanche, c'est sur la finalité de ces IA que la différence se jouera car elle implique le consentement explicite de l'utilisateur final. »

Et sur ce point, l'éthique s'impose comme un levier de compétitivité majeur face aux GAFA et autres startups chinoises. Pourquoi ? Parce que si les États-Unis ont emprunté une voie résolument libérale et l'État chinois, celle du contrôle étatique omnipotent, l'Europe, quant à elle, a opté pour une 3e voie différenciante basée sur le respect du citoyen et de l'éthique. Toutefois, pour faire de l'éthique un véritable facteur compétitif, il importe également de bien choisir ses fournisseurs et s'assurer que toute la chaîne de valeur de la donnée est éthique de bout en bout, de la collecte initiale à la livraison du service à l'utilisateur final, en passant par le stockage et le traitement de la data.

L'avis de l'expert

Emmanuel Dubois,
cofondateur d'Indexima

Le RGPD propose un cadre certes contraignant mais qui permet une définition et une utilisation de la donnée. En ce sens, il peut donc alimenter les entreprises et les amener à mieux exploiter leurs big data, notamment face aux GAFA. Voilà quel est le réel enjeu de la donnée et l'importance de la qualité de ces données : une donnée unifiée, cohérente, vérifiée, traçable et exploitable.

Les GAFA et les licornes les plus importantes sont capables à elles seules de disrupter le marché et de faire trembler les plus grandes entreprises mondiales, autrefois indétrônables, et ce, quel que soit le secteur. Désormais, ces entreprises « traditionnelles » savent qu'elles ne sont plus pérennes. En 10 ans, elles n'ont donc eu d'autre choix que de devoir se transformer mais souvent de façon anarchique et hiérarchique. Résultat, on se retrouve avec des organisations très peu agiles et dont la transformation relève surtout de la communication marketing. Si toutes ont pris conscience de l'importance des données, peu parviennent à en faire un levier. La raison : elles se montrent incapables d'exploiter la mine d'informations qu'elles possèdent depuis des dizaines d'années, et ce malgré toute la richesse qu'elles représentent. Là où les GAFA sont *digital natives et customer centric*, et ont mis en place dès le 1er jour des KPI pour piloter leurs activités, réduire en permanence leurs coûts, optimiser leurs process, innover et améliorer sans cesse leur connaissance client. Aujourd'hui encore, l'algorithme d'Amazon reste plus performant que celui proposé par les organisations historiques, à l'image de ce que peut proposer une Fintech par rapport à une banque classique. Une incapacité qui, en cas de nouvelle crise financière, risque d'accélérer la faillite de nombreux établissements.

Toutefois, si les géants du web sont résolument *consumer centric*, ils n'ont souvent que faire de l'éthique… au contraire des entreprises plus « institutionnelles » qui, elles, se montrent plus respectueuses des règlements. C'est là que le RGPD peut amener l'entreprise vers la data éthique. Toutefois, elles ont un vrai devoir d'éducation des utilisateurs quant à leurs droits vis-à-vis de leurs données personnelles. Voilà comment elles vont pouvoir se différencier des entreprises moins éthiques : à travers une transparence totale sur l'utilisation des données et la communication auprès du grand public. Mais si la data est un vrai levier de compétitivité face aux GAFA, elle reste avant tout un sujet d'organisation. N'est pas *data driven* qui veut et l'organisation est clé pour le devenir. Donc repenser son organisation non plus autour des process mais autour de la data est une question d'état d'esprit.

Pas de data (éthique), pas d'IA !

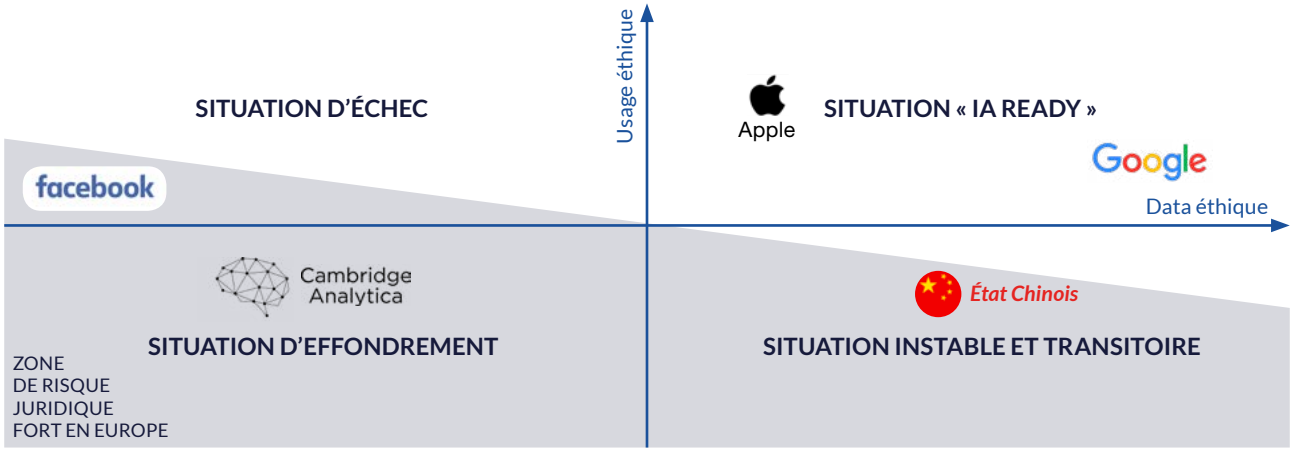
Dans ce contexte de guerre des données et des intelligences, quelle place peut réellement occuper une entreprise française – ou européenne – face à la puissance des GAFA et autres Alibaba ? Une place de premier choix ! Comment ? C'est très simple, parce que « *nous sommes les meilleurs* ». Ce n'est pas nous qui l'affirmons mais Luc Julia lui-même. Et d'enfoncer le clou : « *Nous avons les meilleurs mathématiciens au monde donc nous avons un vrai rôle à jouer. La France a toutes les cartes en main pour devenir le phare de l'intelligence artificielle dans les années à venir… Tout autant que les Chinois et les Américains aussi peuvent l'être.* »

Une entreprise sur quatre

VOIT LA MOITIÉ DE SES PROJETS MENÉS EN INTELLIGENCE ARTIFICIELLE ÉCHOUER[3].

Il n'est possible pour personne de stopper l'avènement de l'IA donc vous n'avez pas d'autre choix que d'être prêts, et ce alors même que la Commission européenne a publié en 2019 ses lignes directrices en matière d'éthique pour le développement d'une intelligence artificielle digne de confiance[4], et la nouvelle directrice de la Commission Européenne, Ursula von der Leyen, a même déclaré qu'un règlement européen autour de l'IA en complément du RGPD, était une de ses priorités. En effet, l'exigence d'immédiateté imposée par les consommateurs va nécessairement faire appel à l'IA demain. L'intelligence artificielle tend à devenir une norme comme le fait d'avoir une adresse email gratuite en son temps. « *Or rien n'est gratuit et les GAFA ne font pas d'humanitaire* », rappelle Cédric Missoffe. En conséquence, les cartes risquent d'être rebattues entre ceux qui l'auront compris et ceux qui ne veulent pas le voir, parmi lesquelles certaines GAFA peut-être ?

Dès lors, les entreprises peuvent être indexées en quatre catégories. Si les GAFA oscillent entre une data et des usages plus ou moins éthiques, le pas qui les sépare d'un scandale n'est pas si grand. Quant aux entreprises chinoises soumises à l'obligation de confier leurs données au gouvernement, elles n'ont aujourd'hui pas vraiment la possibilité de pratiquer une IA éthique malgré des données, qui elles en revanche, sont fiables et de qualité. Pour faire la différence et s'imposer sur le marché, les entreprises françaises et européennes doivent donc s'appuyer sur les règlementations en place pour tendre vers le quadrant en haut à droite et tendre vers une data éthique, condition nécessaire à une intelligence artificielle digne de confiance.



« *L'éthique offre la possibilité aux entreprises de contribuer librement à l'IA mais en respectant certaines règles pour le bien de tous*, conclut Didier Gaultier, directeur Data Science & AI de Business & Decision. *C'est pourquoi Data éthique et IA sont aujourd'hui indissociables sous peine soit de risquer à terme de lourdes pénalités règlementaires, soit à plus brève échéance, de risquer de perdre la confiance de leurs clients et utilisateurs.* »

Conclusion

Qui a déjà pris le temps de lire les conditions générales d'utilisation (CGU) des données d'un site web ? Une étude américaine réalisée en 2016 démontrait que 74 % des personnes validaient les CGU sans même les avoir ouvertes et 98 % ne prenaient pas le temps de les lire[1]. Pourquoi ? Parce qu'elles sont souvent incompréhensibles.

Et sur ce point, le RGPD n'a malheureusement rien changé ! **Fin 2018, seuls 1 % des sites internet étaient capables de répondre dans les 30 jours aux requêtes emails concernant les données personnelles en leur possession**[2]. 83% n'étaient pas en mesure de répondre et dans 16 % des cas, l'adresse email mentionnée s'avérait invalide. *« On peut regretter l'absence d'une éthique globale ou générique acceptée par tous, constate Ada Sekirin, directrice International de Business & Decision. Or, il est indispensable de pouvoir dire à l'algorithme ce qu'il doit faire dans un cadre X, Y ou Z acceptable par tout le monde. Il faut donc prendre une position et c'est très compliqué. Ces blocages sont encore fortement présents et le débat doit se situer à un niveau plus fondamental que le niveau actuel. »*

Dès lors qu'une data est non éthique, elle cesse d'être fiable, et elle induit par conséquent potentiellement un grand nombre d'erreurs tout au long de la chaîne de processus associés : dans son traitement, dans son analyse, dans sa restitution, dans son utilisation, y compris dans le raisonnement des concepteurs et donc, de fait, dans le processus d'apprentissage des machines.
« Une simple donnée fausse va, au travers de l'IA, être croisée avec des centaines d'autres données, qui, même si elles sont toutes justes, donnera au final des informations fausses, martèle Didier Gaultier. *Et entraîner une IA avec des informations fausses ou incomplètes représente un très gros risque. Une donnée clé manquante sur l'environnement d'une IA peut également avoir des conséquences imprévisibles. »*
En septembre dernier, le constructeur automobile Tesla testait ainsi sa nouvelle fonction semi-autonome « Smart Summon » sur un de ses Model X. La voiture était censée sortir seule de sa place de parking pour rejoindre son « propriétaire » grâce à la géolocalisation de son smartphone. Mais c'était sans compter la réaction des individus « lambda » face à un véhicule sans conducteur dans un parking, lesquels se sont mis à hésiter et cela a même créé quelques situations de blocage. « Perturbée », la voiture autonome a alors perdu le contrôle et causé certains accidents heureusement sans gravité.

« Le problème se situe bien sûr au niveau du raisonnement placé dans l'IA relatif à l'analyse de l'environnement, mais également dans tout le processus d'apprentissage, explique Didier Gaultier, directeur Data Science & AI de Business & Decision. *Pourquoi ? Parce que les individus ont eu peur en voyant ces voitures sans pilote et ont hésité, adoptant alors un comportement inhabituel pour lequel la "machine" autonome n'a pas été entraînée. Le facteur émotionnel des autres usagers et les complexités de l'environnement global d'un parking n'ont pas été suffisamment pris en compte. »*

On voit donc que pour qu'une IA fonctionne correctement, il faut qu'elle s'appuie non seulement sur une analyse correcte de son environnement, mais avant toute chose sur des données complètes, correctes, cohérentes, intègres, non biaisées et de qualité ; en un mot : éthiques.
Une règle de base est qu'une IA ne peut jamais être meilleure ou plus performante que ce qu'elle a déjà appris[3], l'apprentissage sur des données non éthiques ne permettra donc en aucun cas de créer une IA éthique et fiable. Une IA non fiable, on l'a vu précédemment au travers de nombreux exemples, va obligatoirement créer un « bad buzz », et ne manquera pas de dégrader l'image et la réputation.

C'est pourquoi, une IA digne de confiance ne peut et ne doit par conséquent s'appuyer que sur une data éthique.

Notes

PARTIE 1

1 Étude Données personnelles et confiance : évolution des perceptions et usages post-RGPD, réalisée par la Chaire Valeurs et Politiques des Informations Personnelles de l'Institut Mines-Télécom et Médiamétrie, octobre 2019

2 ConsoGlobe – https://www.planetoscope.com/Internet-/1523-informations-publiees-dans-le-monde-sur-le-net-en-gigaoctets-.html

3 Baromètre 2018 du CESIN – https://www.silkhom.com/cybersecurite-3-barometres-des-cyberattaques-a-connaitre-en-2019/

4 Source : Data Security Breach, 2017

5 Selon le Breach Level Index

6 Rapport F-Secure, 2019 – https://fr.press.f-secure.com/2019/03/05/cyber-attaques-h2-2018/

7 Étude Accenture 2019 sur la cyber-résilience – https://www.accenture.com/fr-fr/insights/security/invest-cyber-resilience

8 Rapport McAfee en partenariat avec le Center of Strategic and International Studies, février 2018 – https://siecledigital.fr/2018/02/22/600-milliards-de-dollars-cest-le-cout-de-la-cybercriminalite/

9 Rapport « Data Trust Readiness », réalisée par Talend et Opinion Matters en avril 2019 - https://info.talend.com/datatrustreadinessreportfr.html

10 Données personnelles et confiance : évolution des perceptions et usages post-RGPD, étude IMT-Médiamétrie, 2019

11 Étude du Pew Research Center, 2018 – https://www.pewresearch.org/fact-tank/2018/09/05/americans-are-changing-their-relationship-with-facebook/

12 Au 3e trimestre 2019, https://www.journaldunet.com/ebusiness/le-net/1125265-nombre-d-utilisateurs-de-facebook-dans-le-monde/

13 Données personnelles et confiance : évolution des perceptions et usages post-RGPD, étude IMT-Médiamétrie, 2019

14 Source : L'usine digitale, 2016 : https://www.usine-digitale.fr/article/donnees-personnelles-3-4-des-francais-ne-font-pas-confiance-aux-applis-mobiles.N387806

15 Données personnelles et confiance : évolution des perceptions et usages post-RGPD, étude IMT-Médiamétrie, 2019

16 Étude Talend, 2018 – https://fr.talend.com/about-us/press-releases/the-majority-of-businesses-are-failing-to-comply-with-gdpr-according-to-new-talend-research/

17 Source : KPMG International, « Crossing the line – Staying on the right side of consumer privacy », 2016 : https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2016/11/crossing-the-line.pdf

18 Orange, octobre 2017 – https://hellofuture.orange.com/fr/chiffrement-homomorphe-la-cle-de-la-securite/

PARTIE 2

1 Données personnelles et confiance : évolution des perceptions et usages post-RGPD, étude IMT-Médiamétrie, 2019

2 Selon le Baromètre de la Confiance des Français dans le numérique, ACSEL, 2017 – https://www.acsel.eu/presentation-de-6eme-vague-barometre-de-confiance-francais-numerique/

3 D'après une étude de la société Norton Lifelock, 2019

4 Données personnelles et confiance : évolution des perceptions et usages post-RGPD, étude IMT-Médiamétrie, 2019

5 Selon une étude réalisée par Experian marketing Services, 2017 – https://www.experianplc.com/media/news/2017/nouveau-livre-blanc-donn%C3%A9es-et-entreprises-en-2017-un-diagnostic-complet/

6 Ontologie : En informatique et en science de l'information, une **ontologie** est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l'« ontologie est aux données ce que la grammaire est au langage » (Wikipédia)

7 Informatique News, juin 2019 – https://www.informatiquenews.fr/rgpd-plus-de-95-000-plaintes-deposees-les-amendes-commencent-a-tomber-62104

8 Femmes du numérique, avril 2018 – https://femmes-numerique.fr/quelle-place-pour-les-femmes-dans-le-numerique/

9 Baromètre de l'innovation signé Odoxa-Microsoft-L'Usine Digitale, 2016 – https://www.usine-digitale.fr/article/donnees-personnelles-3-4-des-francais-ne-font-pas-confiance-aux-applis-mobiles.N387806

10 Selon une étude de l'agence Fuse, 2018 – https://www.businesswire.com/news/home/20180628006409/en/Teens%E2%80%99-Views-Social-Activism-Marketing-Matters-Brands/

PARTIE 3

1 Codeur Mag, 2018 – https://www.codeur.com/blog/avis-clients-importants/

2 Étude du Capgemini Research Institute, 2019 – https://www.decideo.fr/Plus-d-un-an-apre%CC%80s-l-entree-en-vigueur-du-RGPD-seul-28-des-entreprises-declarent-e%CC%82tre-en-conformite_a11357.html

3 Selon une étude réalisée début 2019 par IDC auprès de quelque 2 500 organisations à travers le monde.

4 https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

CONCLUSION

1 https://www.numerama.com/politique/182421-mettez-ce-que-vous-voulez-dans-les-cgu-on-accepte-nimporte-quoi.html

2 Enquête menée par Freebip, décembre 2018 – https://comarketing-news.fr/rgpd-6-mois-apres-le-constat-est-accablant/

3 Les différents modes d'apprentissage de l'IA et leur mise en pratique seront traités dans un livre blanc à paraitre.

Remerciements

Didier Gaultier
Directeur DataScience & AI - Business & Decision

-

Faycal Boujemaa
Head of Strategy - Orange

Emmanuel Dubois
Président et cofondateur - Indexima

Jean-Michel Franco
Product Marketing Senior Director - Talend

Luc Julia
CTO & Senior Vice President of Innovation - Samsung

Business & Decision :

Romain Bernard
Manager DataScience Paris-Nord Ile -de-France

Serge Blanc
Manager Data Science

Michaël Deheneffe
Directeur de la stratégie et de l'innovation

Jérome Dewever
Manager Conseil Expertise MDM

Maxence Dhellemmes
Directeur R&D

Valérie Lafdal
Directrice Générale

Mick Levy
Directeur de l'Innovation Business

Cédric Missoffe
Directeur de l'agence Conseil & Expertise

Ada Sekirin
Directrice région Bénélux et international

Mondher Sendi
Expert AI - Computer Scientist

Stéphane Walter
Manager Conseil & Expertise / Big Data

Business & Decision
Cœur Défense A , 110 Esplanade Général de Gaulle,
92931 Paris La Défense Cedex
www.businessdecision.com
blog.businessdecision.com

Business & Decision - RCS B 384 518 114 - Réalisation : L'agence Bespoke - Crédits photos : iStock

Business & Decision