ia digne de confiance

—CONSTRUIRE UNE GOUVERNANCE
ADAPTÉE À CHAQUE ENTREPRISE

IMPACT AI

— PRÉFACE

La France a l'ambition de devenir un acteur international majeur de l'intelligence artificielle (IA) et voici deux ans que le Gouvernement a défini la stratégie nationale qui la matérialise. Son développement s'intègre aux propres orientations de la Commission européenne, qui visent en particulier à bâtir une IA digne de confiance et plus souveraine pour l'Europe. Les entreprises, essentielles à la réussite de nos objectifs économiques, doivent également être conscientes des enjeux éthiques et sociaux de ces technologies inédites et assumer leur part de responsabilité.
C'est pourquoi je soutiens les initiatives vertueuses d'Impact AI. Ce guide, fruit de la réflexion d'une quinzaine d'entreprises et organisations nationales, participe de l'action pour nous engager collectivement sur la voie d'une IA digne de confiance.

Renaud Vedel
— COORDONNATEUR DE LA STRATÉGIE NATIONALE POUR L'INTELLIGENCE ARTIFICIELLE

Intelligence artificielle responsable : des paroles aux actes

AVANT

PROPOS

LE DOUTE N'EST PAS PERMIS : EN CETTE PÉRIODE SI PARTICULIÈRE, L'INTELLIGENCE ARTIFICIELLE (IA) RESPONSABLE EST LE SUJET INCONTOURNABLE. Hier réservé à un public de spécialistes ou de techniciens, il a désormais investi notre conscience collective, et s'est imposé comme l'un des grands enjeux auxquels nous devons réfléchir ensemble. Il est à présent au cœur de la conversation. Il nous rassemble autour de la table et nous interroge sur le chemin que nous voulons tracer collectivement.

Même si aujourd'hui nous n'avons pas toutes les réponses techniques pour garantir une IA sans défaut, son utilisation et sa valeur sont une réalité. Aussi bien les entreprises que les institutions publiques doivent continuer à investir dans la recherche pour développer une IA digne de confiance. En parallèle, nous observons une authentique volonté, de la part de grandes et moyennes entreprises conscientes de leur impact dans la société et de la part des startups cherchant à construire un monde meilleur, d'implémenter des mécanismes pour nous permettre de mieux gérer cette nouvelle technologie de manière responsable.

Témoins de ce mouvement, les organisations publiques et privées, quelles que soient leur taille et leurs missions, sont de plus en plus nombreuses à s'investir dans cette réflexion. Elles s'interrogent alors sur l'usage qu'elles font de l'intelligence artificielle et remettent en question leurs pratiques. Elles mesurent l'impact des décisions et recherchent le point d'équilibre : comment accélérer l'adoption de l'IA tout en respectant les principes éthiques de nos sociétés ? Les dirigeants connaissent l'importance des orientations stratégiques prises aujourd'hui et savent que ces décisions impliqueront des conséquences à long terme.

C'est à travers des initiatives telles qu'Impact AI que ces sujets ont émergé dans notre écosystème français et que la réflexion sur l'IA digne de confiance continue à se renforcer jour après jour pour tendre vers une réalité opérationnelle. Aujourd'hui, chacun peut le constater, nous avons quitté la théorie pour entrer dans le champ de la pratique et de l'exécution. C'est pourquoi il est important de traduire les grands principes de l'IA digne de confiance pour en faire émerger des orientations fortes et des solutions concrètes.

C'est dans ce sens que le collectif Impact AI a axé son travail pour aboutir à ce Guide pratique. Notre ambition est de parvenir à construire des outils que chaque organisation pourra s'approprier pour alimenter son propre questionnement sur des thèmes tels que la dignité humaine, la transparence ou encore l'équité et la non-discrimination. Qu'est-ce qu'une IA digne de confiance ? Comment l'implémenter en entreprise et garantir qu'elle respecte les principes éthiques fondamentaux dans la durée ? Ce sont les questions que nous avons explorées ensemble et sur lesquelles nous espérons faire avancer la réflexion et apporter une partie des réponses.

Cette mission est indispensable et la poursuite de l'effort est vitale.

J'en ai la conviction : plus que jamais, cette mission est indispensable et la poursuite de l'effort est vitale. C'est la raison pour laquelle nous travaillons pour permettre à chacun d'aborder le sujet de l'intelligence artificielle de façon à la fois responsable et opérationnelle. C'est en nous appuyant sur la force du collectif que nous pourrons atteindre cet objectif.

Le document que vous tenez entre les mains représente le fruit d'un travail collectif, qui s'est construit et nourri grâce à un écosystème d'entreprises, de startups et d'institutions académiques. C'est le résultat de l'engagement des parties prenantes qui composent ce formidable collectif et l'enrichissent de leurs particularités. Au nom de tous les contributeurs, je suis fier de vous le présenter.

Bonne lecture.

Marcin Detyniecki
— VICE-PRÉSIDENT D'IMPACT AI
HEAD OF RESEARCH
AND DEVELOPMENT & GROUP
CHIEF DATA SCIENTIST
AT AXA

L'intelligence artificielle digne de confiance est le sujet incontournable.

HIER RÉSERVÉ À UN PUBLIC DE SPÉCIALISTES OU DE TECHNICIENS, IL A DÉSORMAIS INVESTI NOTRE CONSCIENCE COLLECTIVE, ET S'EST IMPOSÉ COMME L'UN DES GRANDS ENJEUX AUXQUELS NOUS DEVONS RÉFLÉCHIR ENSEMBLE.

IA digne de confiance : un avantage compétitif durable pour les organisations françaises

EDITO

GESTION DES INFRASTRUCTURES IT, ÉCOUTE CLIENT, AMÉLIORATION DU PARCOURS COLLABORATEUR, L'INTELLIGENCE ARTIFICIELLE (IA) CONTINUE D'ÊTRE DÉPLOYÉE DANS TOUTES LES DIMENSIONS DE L'ORGANISATION. Si la valeur de l'IA n'est plus à démontrer, l'usage de cette technologie n'est pas sans risques. Prenons l'exemple des modèles comportementaux (i.e. ciblage marketing consommation), s'ils n'étaient pas correctement calibrés, ceux-ci pourraient être sujets à des biais potentiellement discriminants pour le consommateur. L'interrogation serait similaire sur d'autres catégories de modèles (ex: reconnaissance faciale ou traitement du langage naturel), si ces derniers étaient développés par des ressources non sensibilisées aux risques IA, à partir des données biaisées et sans mécanisme de contrôle. C'est dans ce contexte que les principales autorités de régulation ont initié des projets de réglementation de l'IA, en commençant par la Commission Européenne.

Au-delà des exigences réglementaires, prendre en compte l'éthique et initier une gouvernance autour de l'IA devra s'imposer comme un prérequis pour toute organisation développant ou utilisant cette même technologie. Remarquons cependant que plus de 60 % de ces organisations s'estiment insuffisamment prêtes pour répondre à ces nouveaux risques[1].

Au regard de ces enjeux, nous proposions un cycle de travaux sur la gouvernance de l'IA aux membres d'Impact AI dès février 2020, ayant pour objectif de recenser et d'identifier les besoins des organisations. Quels seraient les besoins de gouvernance des organisations sans cadre réglementaire défini ? Comment traduire des principes de robustesse ou de transparence, dans les opérations, autour de l'IA ?

1. Deloitte (2020). Thriving in the era of pervasive AI, Deloitte's State of AI in the Enterprise, 3rd Edition.

RIAL

Quelles seraient les bonnes pratiques de gouvernance de l'IA ? C'est l'essentiel de ces réflexions et de nos enseignements auprès d'acteurs majeurs de l'IA en France que nous vous proposons de partager à travers ce guide pratique.

C'est à travers ce groupe de travail que nous avons pu recenser les initiatives de nombreuses organisations autour de leur usage responsable de l'IA (réflexions, travaux, ressources dédiées, mise en place de processus…) afin d'identifier, par domaines, des zones de risques et des actions de remédiation. S'il peut être facile d'obtenir des résultats probants avec l'IA, les pièges sont nombreux !

Prendre en compte l'éthique et initier une gouvernance autour de l'IA devra s'imposer comme un prérequis.

Gageons que ce guide vous éclairera sur les enjeux et risques de l'IA pour sensibiliser votre organisation ; qu'il vous donnera aussi quelques bonnes pratiques clés pour initier la mise en œuvre d'une gouvernance efficiente de vos projets d'IA.

N'en doutons pas, les enjeux règlementaires émergents impacteront les organisations. Considérons-les comme une opportunité unique de replacer l'humain au cœur des décisions, à travers la déclinaison d'une gouvernance forte autour de l'usage de l'IA. Comme à travers une IA digne de confiance, robuste et explicable. De par ces initiatives, les précurseurs sauront maintenir leur avantage concurrentiel – car suffisamment matures dans la remédiation de ces nouveaux risques, tout comme dans l'industrialisation de leurs projets d'IA.

Fruit d'un engagement et de convictions partagés, nous vous partageons ces travaux à travers ce guide pour que vous puissiez bénéficier de l'IA, de manière responsable et éthique. Expérimentez les initiatives que nous restituons ici ! Et rejoignez l'écosystème d'Impact AI avec vos expériences de gouvernance et de mise en œuvre d'une IA responsable dans vos organisations. Nous vous y convions chaleureusement !

Bonne lecture.

Grégory Abisror
— ÉLU AU CONSEIL D'ADMINISTRATION D'IMPACT AI
ASSOCIÉ DATA & AI
DE DELOITTE FRANCE

N’en doutons pas, les enjeux règlementaires émergents impacteront les organisations.

CONSIDÉRONS-LES COMME UNE OPPORTUNITÉ UNIQUE DE REPLACER L’HUMAIN AU CŒUR DES DÉCISIONS.

Qu'est-ce qu'Impact AI ?

Créé en 2018, le collectif de réflexion et d'action Impact AI ambitionne d'éclairer les enjeux éthiques et sociétaux de l'intelligence artificielle et de soutenir des projets innovants et positifs pour le monde de demain.
Constitué de grandes entreprises, d'ESN (Entreprises de services numériques), de sociétés de conseil en stratégie, de startups et d'écoles, il a vocation à travailler avec tout l'écosystème numérique. Il associe à ses initiatives les acteurs économiques, les institutions, les organismes de recherche et les représentants de la société civile, pour créer une approche de l'IA responsable et inclusive qui réponde aux besoins et attentes des citoyens.

Face à l'ampleur des enjeux, Impact AI veut, à travers ses productions (études, échanges, événements, rencontres, baromètre, livres blancs), prendre de la hauteur, nourrir le débat public, stimuler et accompagner l'émergence d'une IA digne de confiance ainsi que formuler des recommandations à l'intention des pouvoirs publics. Ce collectif promeut également le développement et le partage d'outils favorisant et vérifiant l'usage responsable de l'intelligence artificielle. Enfin, il encourage et soutient des projets innovants dans tous les secteurs économiques qui œuvrent dans l'intérêt général.

Qu'est-ce que le guide pratique « IA digne de confiance : construire une gouvernance adaptée à chaque entreprise » ?

Ce document retrace le travail d'une quinzaine d'entreprises françaises (voir p. 94) rassemblées au sein du groupe de travail IA Responsable d'Impact AI. En pleine pandémie de la Covid-19, ces organisations ont continué à se réunir virtuellement et en toute confidentialité afin d'explorer et d'analyser leurs enjeux de la gouvernance de l'intelligence artificielle. Chaque partie résume ainsi les discussions des ateliers en ligne effectués entre mars et juin 2020 afin de :

Partager des expériences de mise en œuvre des principes éthiques avec les autorités travaillant sur des projets de réglementation ;

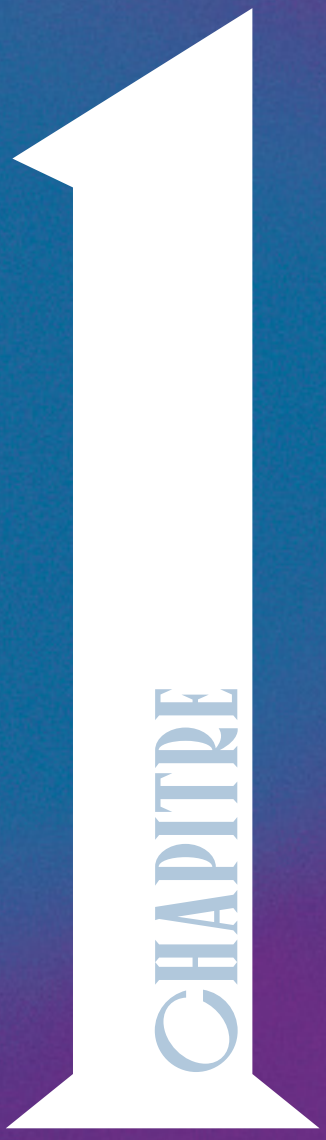
Présenter des bonnes pratiques afin de soutenir les entreprises qui s'engagent sur la voie de la gouvernance de l'IA ;

Sensibiliser plus largement d'autres organisations et entreprises aux enjeux de l'intelligence artificielle digne de confiance.

De l'IA Responsable à l'IA digne de confiance

À la suite de nombreux débats au sein du groupe de travail, le guide pratique a souhaité se concentrer sur le concept d'IA digne de confiance car il permet d'embrasser les dimensions éthiques, techniques et légales de l'Intelligence Artificielle (voir glossaire p. 96), tandis que le concept d'IA Responsable traite en priorité des enjeux de responsabilité des entreprises. D'autre part, la distinction permise par l'IA digne de confiance est très appréciée à l'heure où les organisations se concentrent sur l'acceptation et l'adoption des applications d'intelligence artificielle par les utilisateurs.

— SOMMAIRE


CHAPITRE

Éthique et intelligence artificielle : des enjeux économiques, sociétaux et normatifs

Quels sont les enjeux éthiques de l'intelligence artificielle (IA) ? Impact AI a analysé les risques spécifiques intrinsèques de cette technologie puis les conditions extrinsèques de sa généralisation, telles que l'acceptabilité sociale et l'encadrement juridique et réglementaire. À la demande des autorités européennes, le collectif a pu ainsi formuler des recommandations pratiques, capables de soutenir les organisations désireuses de mettre en place une IA digne de confiance.

LE DYNAMISME DE L'IA SE CONFIRME EN FRANCE. D'après l'étude de la DGE parue en février 2019, « Intelligence artificielle : état de l'art et perspectives pour la France[1] », elle se diffuse dans de nombreux secteurs d'activité, impacte l'économie et fait l'objet de stratégies territoriales de déploiement. Au total, les analyses prospectives démontrent qu'elle représente de véritables opportunités.

Durant la crise de la Covid-19, en 2020, l'IA devient un outil clé pour lutter contre la pandémie et garantir la continuité des activités des entreprises françaises. D'une part, les experts scientifiques en intelligence artificielle ont considérablement aidé les recherches médicales des organismes internationaux[2] et ont prêté main forte aux soignants sur le terrain[3]. Plus largement, le confinement a conduit les entreprises françaises à faire basculer en un temps record leur organisation dans le télétravail, preuve que la maturité technologique ne constitue plus une option mais un impératif. Et le développement de l'intelligence artificielle renforcera cette nécessité. Cela souligne la dépendance au numérique des organisations et leur fragilité en cas de failles dans les systèmes, notamment de cybersécurité, comme en témoigne l'attaque de la HHS (administration américaine chargée de la politique de la santé) au paroxysme de la crise[4].

Le gouvernement a lancé un plan national sur l'IA en 2018 et annoncé un investissement de 1,5 milliard d'euros à l'horizon 2022. Sa stratégie s'appuie sur trois piliers : le renforcement de l'offre française, la diffusion technologique dans les PME avec le soutien des grands groupes, la construction d'une IA transparente et respectueuse des libertés publiques.

Occupant la 15e place dans l'indice *Digital Economy and Society*[5] de la Commission européenne, la France atteint à peine la moyenne européenne. Les membres du collectif Impact AI veulent concourir à la réussite de la feuille de route

1. Atawao consulting et al., « Intelligence artificielle: état de l'art et perspectives pour la France : rapport final », 2019.

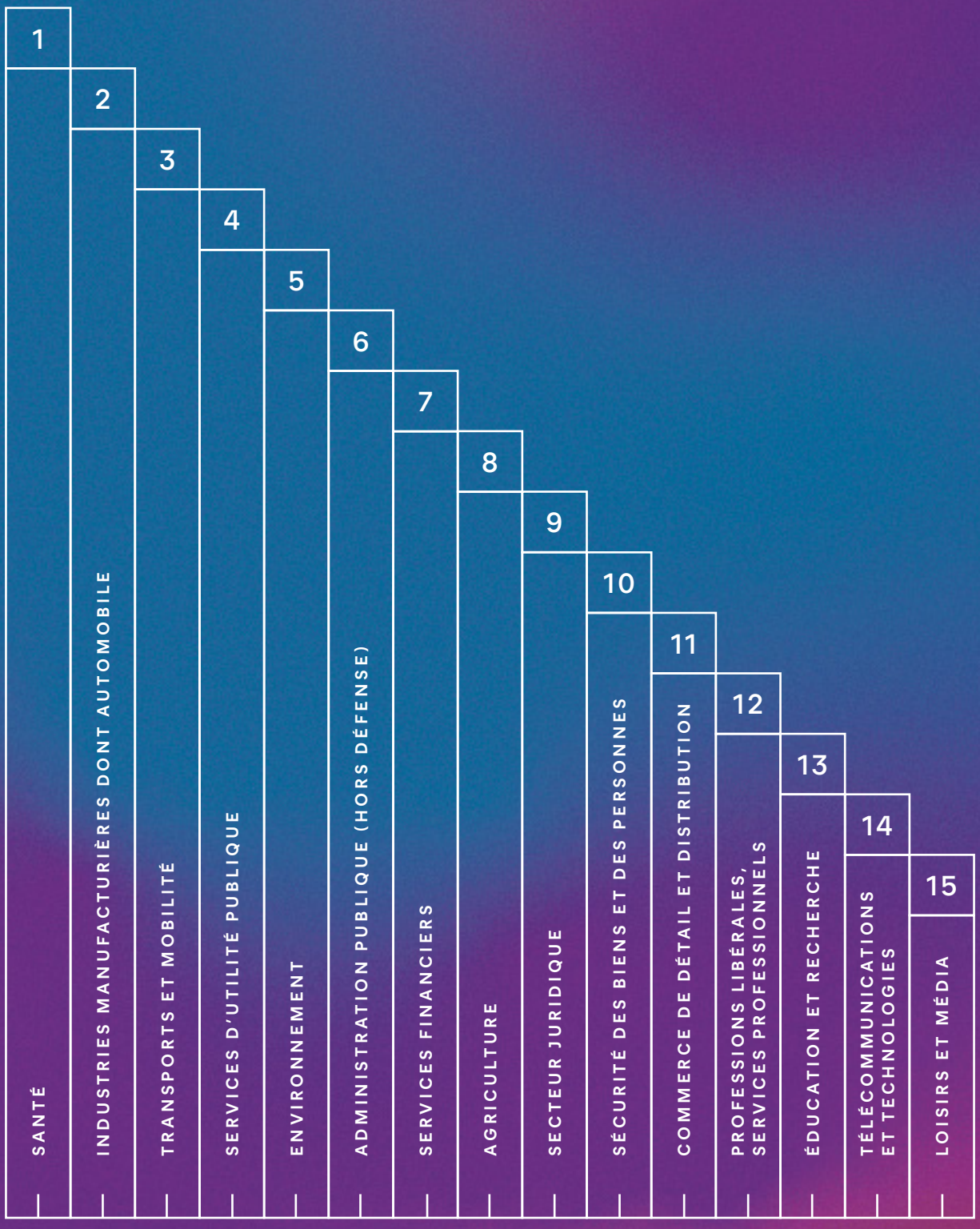
2. Larousserie, David. « Coronavirus : comment l'intelligence artificielle est utilisée contre le Covid-19 », *Le Monde.fr,* 20 mai 2020.

3. Inria. « Inria présente les projets de sa mission covid », *inria.fr,* mai 2020.

4. Stein, Shira, Jennifer Jacobs. « Cyber attacks hits U.S. Health Agency Amod Covid-19 outbreack », *Bloomberg.com*, mars 2020.

5. « The Digital Economy and Society Index (DESI) », Shaping Europe's digital future - European Commission, février 2015.

<fig.1.> L'étude de la DGE de 2019 montre la variété des secteurs d'activités qui seront impactés par l'intelligence artificielle en France dans les prochaines années.



des autorités nationales afin de transformer le pays en un acteur majeur de l'intelligence artificielle. Considérant la confiance des utilisateurs dans cette technologie comme une condition sine qua non de sa réussite, ils se donnent ainsi la mission de diffuser largement les principes et pratiques d'une IA éthique et responsable.

Découverte des enjeux de l'IA digne de confiance

Qu'est-ce-que l'intelligence artificielle ? La réponse à cette question n'a rien d'évident selon que l'on se focalise sur tel ou tel de ses aspects technologiques. D'aucuns la considèrent comme une technique d'apprentissage automatique à partir de données massives (big data), d'autres comme capable de comportements imitant ou ressemblant à ceux des hommes. Quels qu'en soient ses contours et l'ensemble des facteurs et fonctions mathématiques dont elle découle, son résultat demeure identique : l'IA effectue des calculs hors de portée de n'importe quel être humain, qui permettent de mieux analyser des situations et de prendre, statistiquement, de meilleures décisions. Ce constat pose dès lors une question cruciale : quelle confiance accorder aux résultats qu'elle produit ?

— DES RISQUES SPÉCIFIQUES, MAIS PAS SEULEMENT

Les bienfaits des applications de l'IA s'accompagnent d'une part indéniable de risques. La technologie peut en effet être utilisée pour concevoir des modèles prédictifs — par exemple des comportements — et permettre une prise de décision automatisée ce qui peut engendrer des problèmes complexes et multiples[6]. Les exemples ne manquent pas et dans des domaines extrêmement variés : la justice (jugements automatiques discriminatoires), la finance et l'économie (algorithmes boursiers, octroi de prêts bancaires), l'éducation (attribution de places universitaires, orientation vers des cursus), le recrutement (proposition d'offres d'emploi défavorables aux femmes), le contrôle voire la suppression de contenus web en contradiction avec le principe de la liberté d'expression, la sécurité (lutte contre le terrorisme, surveillance et contrôle des foules, reconnaissance faciale…), le marketing et la publicité en ligne (profilage et messages commerciaux personnalisés), la mobilité et la liberté de circulation (véhicule autonome, drone), l'urbanisme (ville intelligente), l'organisation du travail et le marché de l'emploi (assistants virtuels, usine du futur…). Cette énumération illustre que sous l'apparente neutralité des calculs mathématiques des algorithmes, les systèmes d'IA peuvent commettre des erreurs jusqu'à causer des injustices et discriminations répétées[7].

On distingue trois types de causes à cela. D'abord, les anomalies dans les programmes ou dans les données utilisées. Deuxième problème potentiel : un apprentissage automatique fondé sur des données partielles et partiales. Enfin, des choix erronés ou biaisés lors de la conception des algorithmes, conduisant à des résultats inappropriés, discriminatoires par exemple.

Certaines de ces difficultés ne sont pas spécifiques à l'intelligence artificielle. Ainsi, les anomalies ou bugs de programmes informatiques « classiques » sont courants ; les biais se retrouvent aussi dans des logiciels usuels parce qu'ils sont conçus par des êtres humains qui en sont empreints par nature. Les innovations qui font intervenir dans leur fabrication plusieurs types d'acteurs, non formés ou sensibilisés aux risques de l'IA, démultiplient les risques à chaque étape de production. C'est le cas des robots par exemple, où les divers fournisseurs – matériel, équipements… – peuvent introduire involontairement des dysfonctionnements dans la chaîne de valeur.

L'intelligence artificielle reste cependant singulière à deux égards : elle peut reproduire et amplifier des biais existants de manière opaque[8]. Il est donc nécessaire d'édifier des garde-fous afin de se prémunir contre ces dangers.

6. « Algorithmes : contrôle des biais S.V.P. », Institut Montaigne, mars 2020.

7. Par exemple, une publication récente montre que les systèmes de reconnaissance faciale présentent des biais relatifs au genre et à la couleur de peau. Si l'on est femme et noire aux États-Unis, le taux d'erreur peut s'élever à 35 %. (*Sciences et Avenir,* 6 mars 2018)

8. *Op. cit.,* « Algorithmes : contrôle des biais S.V.P. »

L'intelligence artificielle peut reproduire et amplifier des biais existants de manière opaque.

IL EST DONC NÉCESSAIRE D'ÉDIFIER DES GARDE-FOUS AFIN DE SE PRÉMUNIR CONTRE CES DANGERS.

— POUR UNE ACCEPTATION SOCIALE DES PROMESSES DE L'IA

Xavier Guchet qui consacre ses travaux d'épistémologie à l'éthique des technologies contemporaines, définit la science responsable. Il s'agit, explique-t-il, *« d'une science capable d'anticiper les impacts de ses applications techniques sur la nature et sur la société, de façon à opter en amont pour un design technologique qui maximise les bénéfices et minimise les risques encourus, voire les effets dommageables*[9] *»*.

Ce concept s'applique mot-pour-mot à l'intelligence artificielle. Créer une IA responsable donne l'obligation aux acteurs qui proposent des solutions basées sur cette technologie de tenir compte de leurs conséquences sur notre environnement, positives ou non, dès le stade de la conception. Cette démarche, nommée éthique *by design*, consiste concrètement à sensibiliser en amont les concepteurs de services d'intelligence artificielle à l'éthique, que ce soit dans les usages, les méthodologies, le code, la programmation ou le choix des données[10].

Créer une IA responsable donne l'obligation aux acteurs qui proposent des solutions basées sur cette technologie de tenir compte de leurs conséquences sur notre environnement, positives ou non, dès le stade de la conception.

Opter pour un design qui anticipe, évalue et minimise les risques tout en optimisant les bénéfices, impose d'adopter une approche délibérative et collaborative préalable afin de hiérarchiser des valeurs et de trouver une base de travail consensuelle. L'éthique *by design* répond ainsi à une double problématique : d'une part elle établit des critères d'acceptation ou de préférence sociale ; d'autre part, elle prémunit les acteurs, en cas de défaillance, contre les risques juridiques, financiers, réputationnels… Elle nécessite de se doter de mécanismes de contrôle qui garantissent la qualité des données alimentant les algorithmes. L'analyse d'impact identifie les dangers potentiels et propose des mesures pour y faire face.

Si le risque est l'un des attributs de la société moderne (Beck, 2003)[11] sa perception reste subjective. Ainsi, des cas mineurs de mésusages de l'IA, surtout s'ils sont très médiatisés, sont susceptibles d'attiser les craintes et de provoquer des attitudes de rejet. Or, si avant les années 1970 la société semblait ouverte au progrès technologique, celui-ci suscite aujourd'hui de vives réticences voire des tensions. Les mouvements de contestation liés par exemple au changement climatique — nucléaire, gaspillage alimentaire, pollution numérique, crises sanitaires —, esquissent les traits d'une société de l'émotion, de l'indignation, de l'irrationnel, voire du jusqu'au-boutisme (Aschiéri, 2020)[12].

D'autant que le recours à l'intelligence artificielle conduit souvent à de multiples questionnements légitimes. C'est notamment le cas dans les domaines de la santé et de la sécurité. Les bénéfices de l'intelligence artificielle dans le secteur

9. Guchet, Xavier. « L'éthique des techniques, entre réflexivité et instrumentalisation », Revue francaise d'éthique appliquée, N° 2 (juin 2016): 8-10.

10. Cigref et Cabinet Alain Bensoussan Avocats, « Livre Blanc CIGREF « Gouvernance de l'Intelligence Artificielle dans les entreprises » », septembre 2016, p48.

11. Ulrich Beck, *La société du risque : sur la voie d'une autre modernité* (Flammarion, 2008).

12. Gérard Aschiéri, « Sciences et société : les conditions du dialogue », 2020


Pink

médical sont particulièrement importants : recherche clinique, prothèses, robots compagnons, prévention d'épidémies, réduction d'effets secondaires insoupçonnés, mais aussi assistance aux médecins dans le diagnostic et la définition de protocoles de soin… Ces applications si prometteuses nécessitent toutefois l'utilisation de données massives mais particulièrement sensibles puisqu'elles touchent à l'intime. Des questions de fond se posent, notamment au sujet des organisations situées au dehors du monde de la santé. Comment sécuriser les données ? À quels types d'acteurs y ouvrir l'accès ? Quels sont les mésusages possibles ? Les assureurs pourraient, par exemple, ajuster leurs tarifs et contrats à l'état de santé de chacun, générant ainsi une double peine puisqu'à la maladie s'ajouterait un surcoût financier, accroissant de fait les inégalités jusqu'à, potentiellement, l'exclusion. Et que dire aussi du risque d'une médecine technicisée au point de réduire à sa portion congrue l'échange et l'empathie entre soignant et patient ?

Par ailleurs, pour garantir la sécurité des citoyens, l'intelligence artificielle représente une aide précieuse… qui ne va pas sans risque. Les autorités publiques, en particulier de police, sont en effet amenées à croiser des bases de données jusqu'ici indépendantes pour y déceler des corrélations capables de les mener sur la piste de délinquants. La reconnaissance faciale peut aussi être utile mais également constituer une grave menace pour les libertés. Sans un strict encadrement juridique et technique, l'IA pourrait contribuer à mettre en péril des principes démocratiques par l'avènement d'une société de surveillance généralisée.

Ces deux exemples démontrent que la confiance constitue le préalable indispensable à l'acceptation et à l'adoption de l'intelligence artificielle, une confiance qui non seulement s'acquiert par la pédagogie mais qui s'entretient aussi en adoptant de bonnes conditions d'usage dans la durée.

Malgré tout, à la différence du terme « éthique » qui renvoie à des distinctions socio-culturelles, la « confiance » peut rassembler de nombreuses sensibilités. À cet égard, la notion d'une « IA digne de confiance » telle qu'adoptée par la Commission européenne en 2018, paraît beaucoup plus fédératrice. Elle a montré sa capacité à emporter l'assentiment de tous.

Dans ce contexte, Impact AI estime pour sa part qu'il est inenvisageable que l'intelligence artificielle d'aujourd'hui et de demain oublie ou néglige les grands enjeux sociétaux et environnementaux auxquels nous sommes collectivement confrontés. Plus encore, l'extraordinaire potentiel de cette technologie peut inverser des tendances et réduire les inégalités en renforçant le pouvoir d'agir des organisations sociales et solidaires. Pour que l'IA soit acceptée et désirable, ses progrès doivent bénéficier à tous de façon à bâtir une société plus équitable.

Créer une gouvernance proportionnée aux risques

— LES PROBLÉMATIQUES IDENTIFIÉES PAR IMPACT AI

Les membres du collectif souhaitent apporter leur contribution, grâce à des réponses adaptées au niveau de risque selon les entreprises. En généralisant le recours à l'intelligence artificielle, les entreprises espèrent retirer des avantages concurrentiels, améliorer leur compétitivité et offrir une vaste palette de services à leur clientèle. Plus largement, les technologies et les services fondés sur l'utilisation des données massives sont facteurs de croissance économique, d'innovation technologique et de numérisation. Cependant, si la plupart des entreprises posent un regard lucide sur les problèmes éthiques et juridiques soulevés par l'intelligence artificielle, elles ne disposent pas de toutes les méthodes et ressources nécessaires pour les résoudre.

L'entrée en vigueur du Règlement général sur la protection des données (RGPD) le 25 mai 2018 constitue un premier jalon. Mais l'expansion rapide de l'IA est susceptible de renforcer les stéréotypes et les discriminations, ethniques et de

Les progrès de l'IA doivent bénéficier à tous,

POUR QUE LA TECHNOLOGIE SOIT ACCEPTÉE ET DÉSIRABLE.

genre, risques pointés par de nombreux rapports et publications scientifiques (voir bibliographie, p. 108). C'est pourquoi la question des biais (cognitifs, statistiques, économiques) et de leurs conséquences en termes de fiabilité, de sécurité ou d'équité doit trouver réponse. Les algorithmes ne doivent plus être qualifiés de boîtes noires produisant des résultats qui ne peuvent pas être expliqués. La compréhension des mécanismes des biais et de la manière dont on pourrait anticiper ou limiter leurs effets dommageables alimente les réflexions et les travaux de recherche[13].

La confiance constitue le préalable indispensable à l'acceptation et à l'adoption de l'intelligence artificielle.

Éviter les erreurs et garantir le respect des droits des personnes passe par l'éveil des consciences des utilisateurs et leur capacité à contrôler l'usage qu'ils font de cette technologie. Les organisations et entreprises ont un double rôle à jouer en la matière : d'une part, elles ont à faire connaître et comprendre leurs choix en appliquant les principes de transparence et d'explicabilité ; d'autre part, elles doivent révéler les aléas de l'IA à leurs clients et prospects, sans éroder leur confiance. En dehors de ces considérations humanistes et de respect de la dignité humaine, il convient d'envisager aussi les enjeux économiques pour les organisations privées ou publiques. En effet, une défaillance, des mésusages ou des dérives peuvent entacher l'image d'une entreprise (risque réputationnel), lui faire perdre des contrats ou des parts de marché, voire dissuader de postuler des candidats hautement qualifiés dans un domaine recherché.

Trois exemples peuvent éclairer ces problématiques.

Sur Twitter, l'agent conversationnel Tay, après une phase d'apprentissage, a émis des messages racistes, antisémites, homophobes et sexistes, illustrant les difficultés liées à l'apprentissage à partir de données biaisées. Microsoft a dû stopper son service en 2016 après 24h.

La même année, le premier accident mortel en pilotage automatique de Tesla a relancé la polémique sur la voiture autonome et les questions des responsabilités juridiques et techniques.

Le cas de la solution d'IA d'assistance au recrutement d'Amazon, en 2017, qui s'est avérée discriminatoire, apparaît comme exemplaire des causes d'un dysfonctionnement et du péril pesant sur l'entreprise. L'outil qui se basait sur l'examen des CV des collaborateurs recrutés sur une période de dix ans, essentiellement masculins (biais de genre dans les données), écartait les profils féminins. Leçon de cette expérience malheureuse : de nombreux membres du groupe de travail ont témoigné qu'ils prenaient appui sur la gouvernance de l'IA pour corriger ou atténuer ce type de problèmes grâce à des outils, notamment techniques, et des protocoles.

— UN CONTEXTE RÉGLEMENTAIRE FOISONNANT ET ÉVOLUTIF

Les prémices d'un « droit de l'intelligence artificielle » en France remontent aux années 1990. Toutefois, les bouleversements sociaux et techniques nés de cette technologie dépassent si largement le cadre national que des initiatives se sont multipliées tant au niveau international qu'européen pour élaborer un encadrement juridique de l'IA.

L'Union européenne s'est emparée du sujet à partir de 2018, ce qui s'est traduit par plusieurs publications et des étapes marquantes.

13. Patrice Bertail et al., « Algorithmes : biais, discrimination et équité », 2019.

Les entreprises ne disposent pas de toutes les méthodes et ressources nécessaires

POUR RÉSOUDRE LES ENJEUX SOULEVÉS PAR L'INTELLIGENCE ARTIFICIELLE, MÊME SI LA PLUPART DES ENTREPRISES POSENT UN REGARD LUCIDE SUR LES PROBLÈMES ÉTHIQUES ET JURIDIQUES SOULEVÉS PAR CETTE TECHNOLOGIE.

En février 2018 s'est tenu le premier forum sur les impacts sociaux de l'intelligence artificielle, AI4People, au Parlement Européen. Cet événement a été la genèse de la mobilisation publique et privée sur les enjeux d'éthique et de gouvernance de l'IA. Le « Manuel de droit européen en matière de protection des données », paru en mai 2014, dispose que *« l'intelligence artificielle et la prise de décisions automatisées soulèvent la question de savoir qui est responsable en cas de violations de la vie privée des personnes concernées lorsque la complexité et la quantité de données traitées ne peuvent être déterminées avec certitude. Considérer l'intelligence artificielle et les algorithmes comme des produits soulève la question de la responsabilité personnelle, qui est régie par le RGPD, et celle de la responsabilité du produit, qui ne l'est pas. Il faudrait donc des règles sur la responsabilité afin de combler le vide entre la responsabilité personnelle et la responsabilité du produit pour la robotique et l'intelligence artificielle, notamment les décisions automatisées. »*[14]

En juin 2018, la Commission européenne a mis en place le AI HLEG (*High Level Expert Group on AI*), un groupe de 52 experts indépendants, dont l'objectif général est de soutenir la mise en œuvre de la stratégie européenne en matière d'intelligence artificielle. Celle-ci comprend des recommandations sur l'élaboration de politiques futures et sur les questions éthiques, juridiques et sociétales liées à l'intelligence artificielle, y compris les défis socio-économiques[15].

À l'issue d'une consultation publique, ce groupe a rendu en avril 2019 son rapport, E*thics Guidelines for Trustworthy AI*[16]. Il représente un effort pionnier qui s'adosse à la charte des droits fondamentaux de l'Union européenne pour encadrer l'application de l'intelligence artificielle à différents usages. Le document comprend notamment une grille pilote guidant les organisations dans leur évaluation d'un futur produit ou service basé sur l'IA. Il invite les différentes parties prenantes, en particulier les entreprises, à la tester et à faire part de leurs remarques — une démarche dans laquelle Impact AI s'est grandement investie. En effet, le collectif a produit et communiqué au HLEG un document, *Policy Paper,* synthétisant les observations de ses membres, obtenues à partir de cas d'usage (voir p. 36).

Une liste révisée des critères de la grille pilote du HLEG a été publiée le 17 juillet 2020[17]. Elle comprend sept volets principaux :

<1>. Action et contrôle humain : les systèmes d'IA devraient être les vecteurs de sociétés équitables en se mettant au service de l'humain et des droits fondamentaux, sans restreindre ou dévoyer l'autonomie humaine.

<2>. Robustesse technique et sécurité : une IA digne de confiance nécessite des algorithmes suffisamment sûrs et fiables pour gérer les erreurs ou les incohérences dans toutes les phases du cycle de vie des systèmes d'IA.

<3>. Respect de la vie privée et gouvernance des données : il faut que les citoyens aient la maîtrise totale de leurs données personnelles et que les données les concernant ne soient pas utilisées contre eux à des fins préjudiciables ou discriminatoires.

<4>. Transparence : la traçabilité des systèmes d'IA doit être assurée.

<5>. Diversité, non-discrimination et équité : les systèmes d'IA devraient prendre en compte tout l'éventail des capacités, aptitudes et besoins humains, et leur accessibilité devrait être garantie.

<6>. Bien-être social et environnemental : les systèmes d'IA devraient être utilisés pour soutenir des évolutions sociales positives et renforcer la durabilité et la responsabilité écologique.

<7>. Responsabilité : il convient de mettre en place des mécanismes pour garantir la responsabilité à l'égard des systèmes d'IA et de leurs résultats, et de les soumettre à une obligation de rendre des comptes.

Cette version améliorée du document tient compte d'un grand nombre de remarques émises et remontées par Impact AI lors de la période de consultation, c'est-à-dire jusqu'en novembre 2019. Dans un deuxième temps, la Commission

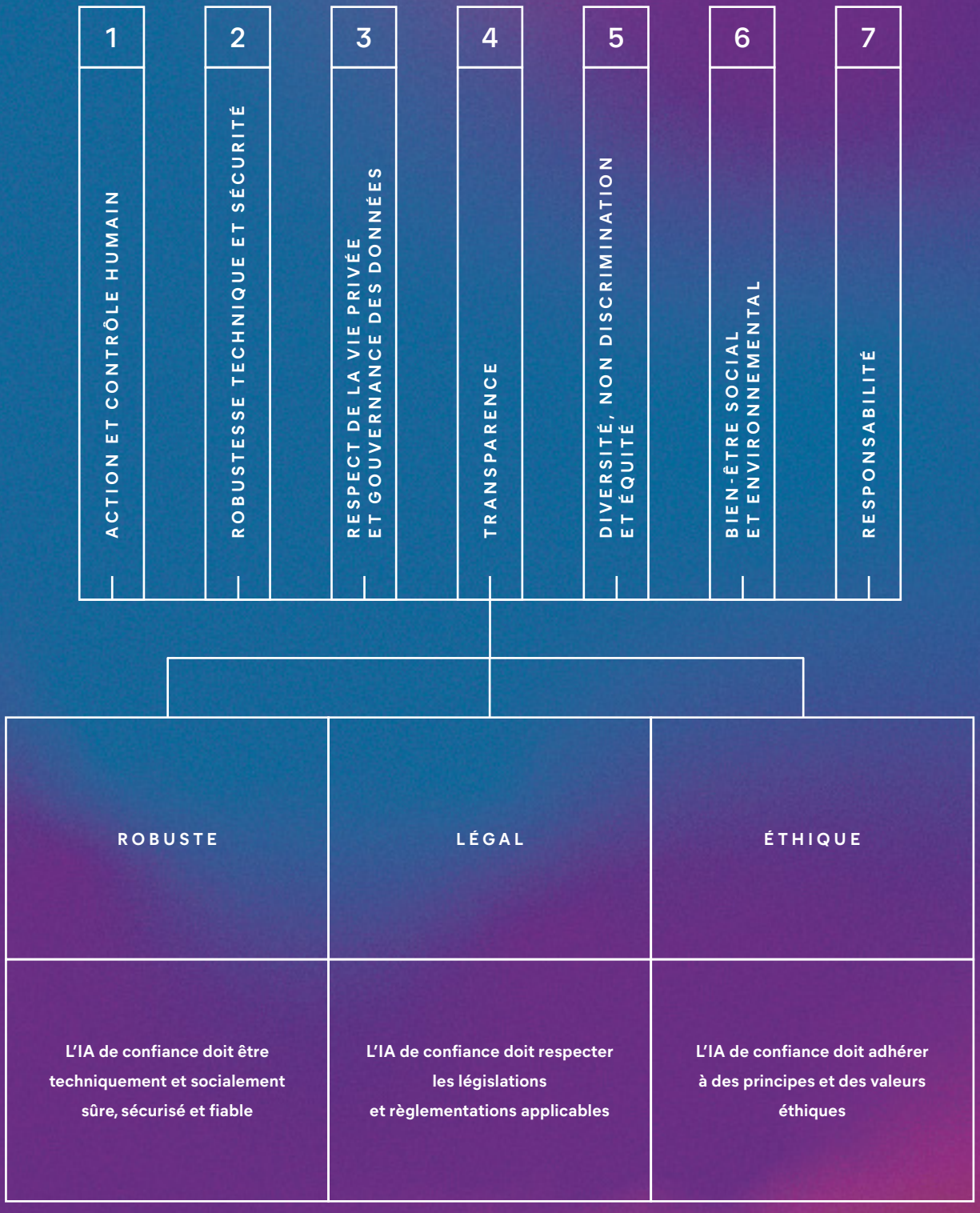
14. « Manuel de droit européen en matière de protection des données », 2018, (pp. 401-402),.

15. « High-Level Expert Group on Artificial Intelligence | Shaping Europe's digital future ».

16. « Ethics Guidelines for Trustworthy AI », Shaping Europe's digital future - European Commission, avril 2019.

17. « Artificial Intelligence: The Commission Welcomes the Opportunities Offered by the Final Assessment List for Trustworthy AI (ALTAI) », Shaping Europe's digital future - European Commission, juillet 2020.

<fig.2.> Les trois piliers de l'IA de confiance se déclinent en sept principes éthiques européens. Ceux-ci deviendraient des prérequis pour tout déploiement de l'intelligence artificielle en Europe.



européenne a publié un livre blanc définissant les grandes orientations stratégiques de l'Union pour promouvoir le déploiement et le développement de l'intelligence artificielle respectant les valeurs communes aux États membres.

Dans ce contexte, le collectif estime qu'il est impératif que les cadres réglementaires de l'intelligence artificielle se concentrent avant tout sur deux objectifs clés : d'abord inciter les acteurs à adopter, volontairement et selon les besoins propres à leur activité, des normes et des procédures de gouvernance leur permettant de mettre en œuvre une IA digne de confiance de façon effective ; ensuite, soutenir le développement de technologies, de systèmes et d'outils qui aident ces acteurs à identifier et atténuer les risques.

> On estime actuellement à plus de 200 les chartes, textes et autres propositions visant à donner un cadre éthique à l'intelligence artificielle, si bien qu'un véritable risque *d'ethical-washing* est apparu pour les institutions ou entreprises désireuses de promouvoir l'IA.

En dépit des difficultés techniques rencontrées et des questions éthiques soulevées, il est indispensable d'atteindre ces objectifs afin de créer les conditions du développement d'une intelligence artificielle fiable, quels que soient les scénarios d'application.

Hors des frontières de l'Union européenne, d'autres initiatives internationales marquantes sont à relever.

Le mouvement *AI for Good* a été lancé en 2017 par l'ONU, en partenariat avec l'ITU (International Telecommunication Union), la XPrize Foundation, l'*Association for Computing Machinery* (ACM) et plusieurs agences des Nations unies (Unesco, Unicef…). Selon ces différentes institutions, les promesses de l'intelligence artificielle pour une vie meilleure pour tous peuvent être tenues, pourvu que les gouvernements, l'industrie et la société civile coopèrent pour un usage positif de la technologie et le contrôle de ses risques.

De nombreuses initiatives ont suivi. En mai 2019 par exemple, les 36 pays membres de l'OCDE, ainsi que l'Argentine, le Brésil, la Colombie, le Costa Rica, le Pérou et la Roumanie, ont adhéré aux « Principes de l'OCDE sur l'intelligence artificielle »[18]. Élaborés par un groupe d'une cinquantaine d'experts de tous horizons, ils prônent le développement d'une IA digne de confiance s'appuyant sur cinq principes fondés sur des valeurs. Ils émettent pour cela cinq recommandations applicables dans le cadre des politiques publiques et de la coopération internationale.

18. « Instruments juridiques de l'OCDE : recommandations du Conseil sur l'intelligence artificielle », 22 mai 2019.

Deux objectifs clés pour les cadres réglementaires de l'IA :

INCITER LES ACTEURS À ADOPTER DES NORMES ET DES PROCÉDURES DE GOUVERNANCE LEUR PERMETTANT DE METTRE EN ŒUVRE UNE IA DIGNE DE CONFIANCE DE FAÇON EFFECTIVE ; SOUTENIR LE DÉVELOPPEMENT DE TECHNOLOGIES, DE SYSTÈMES ET D'OUTILS QUI AIDENT CES ACTEURS À IDENTIFIER ET ATTÉNUER LES RISQUES.

Pour sa part, l'IEEE (*Institute of Electrical and Electronics Engineers*) a lancé une action globale sur l'éthique et l'intelligence artificielle afin de garantir « que chaque partie prenante impliquée dans la conception et le développement de systèmes autonomes et intelligents soit éduquée, formée et habilitée à donner la priorité aux considérations éthiques ». Un premier document cadre de référence a été publié en mars 2019.[19]

De leur côté, l'IEC (*International Electronic Commission*) et l'ISO *(International Standardization Organization*) ont créé un groupe de travail commun, le JTC1/SC42, qui souhaite développer un standard de bonne utilisation des technologies d'IA dans des contextes imposant des normes de sécurité.

Les conférences Asilomar de l'IEEE ont mis en lumière des principes majeurs qui permettent d'appréhender de manière plus claire le champ normatif de l'intelligence artificielle. On estime actuellement à plus de 200 les chartes, textes et autres propositions visant à donner un cadre éthique à l'intelligence artificielle, si bien qu'un véritable risque d'*ethical-washing* est apparu pour les institutions ou entreprises désireuses de promouvoir l'IA.

Compte tenu de ce foisonnement, il est nécessaire de rassembler les différents acteurs de l'écosystème de l'intelligence artificielle afin qu'ils puissent coconstruire le cadre normatif le plus adapté à cette technologie, en tenant compte des différences de problématiques selon les secteurs d'activité. Impact AI propose ainsi dans la partie 2 du guide (voir p. 39, « Des principes généraux à la pratique »), des principes éthiques qui intègrent les initiatives de gouvernance de l'Union européenne, l'OCDE et l'IEEE. Ces principes semblent assez souples pour s'adapter à la majorité des législations nationales. De surcroît, ils se réfèrent à des droits préexistants comme ceux des sciences ou de la médecine, ainsi qu'aux droits humains à la portée universelle.

La mise en place de comités d'éthique s'ouvrant à des personnalités internes ou extérieures à l'organisation apparaît dès lors comme une manière de concrétiser l'application des chartes ou codes d'éthique et d'en faire la promotion en interne auprès des salariés et en externe auprès du public et des clients (voir p. 78-80).

Le groupe de travail IA Responsable, son approche et sa méthode

— GENÈSE DU GROUPE DE TRAVAIL

Souhaitant depuis ses débuts en 2018 bâtir le socle éthique de l'intelligence artificielle, Impact AI a d'emblée créé un groupe de travail IA Responsable qui puisse offrir le fruit de ses réflexions à l'ensemble de l'écosystème. Dans un esprit d'ouverture et collaboratif, sa mission consiste à proposer des solutions concrètes qui inspirent les acteurs, publics et privés, et les épaulent dans leurs efforts de mise en place d'une gouvernance propre à faire advenir un futur de confiance. En somme, il s'agit de prodiguer des conseils pratiques pour aider à construire une intelligence artificielle qui rende le monde à la fois plus ouvert et plus responsable.

Les publications sur ces initiatives et études ont débuté en janvier 2019 à l'occasion de la conférence *Responsible AI*. Puis, les résultats de la première année d'activité et les recommandations d'Impact AI ont été récapitulés dans un livre blanc[20] paru le 8 juillet 2019 pour la conférence annuelle du collectif.

Enfin, des dispositifs techniques et de gouvernance, des formations et des publications ont été regroupés dans une boîte à outils[21], en libre accès sur le site d'Impact AI, qui cherche à faciliter la réflexion et l'action des organisations en la matière. Sans prétendre à l'exhaustivité, elle s'enrichit au fil du temps. Le présent guide s'inscrit dans cette démarche en contribuant à l'effort de formation qui doit être entrepris, comme cela a été largement souligné dans les rapports « Stratégie France IA[22] » (2017) et « Donner un sens à l'intelligence artificielle » (2018) de Cédric Villani[23].

19. IEEE, « Ethically Aligned Design (EAD1e): A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems », 2019.

20. Impact AI, « Un engagement collectif pour un usage responsable de l'intelligence artificielle », 2019.

21. Pour découvrir la boîte à outils d'Impact AI sur l'IA Responsable.

22. France Intelligence Artificielle, « La stratégie IA en France » (ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, mars 2017).

23. Cédric Villani et al., « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », 2018.

— UNE IMPLICATION FORTE POUR RÉPONDRE AUX DEMANDES DE LA COMMISSION EUROPÉENNE

À compter du deuxième semestre 2019 et en 2020, Impact AI a participé activement à l'initiative européenne sur l'intelligence artificielle visant à structurer l'engagement des parties prenantes au sein de l'Alliance européenne de l'IA. Ainsi, en tenant compte des principes d'IA responsable du collectif, trois ateliers et discussions ouvertes ont été organisés dans un double but : d'une part évaluer l'applicabilité du questionnaire AI HLEG alors proposé dans un objectif d'amélioration, d'autre part émettre des commentaires basés sur des études de cas de technologies développées en entreprise, le processus de développement de l'IA et son impact sur la société. Cet exercice, qui a réuni une équipe pluridisciplinaire, s'est avéré très instructif et utile pour les membres du collectif, permettant notamment de confronter des points de vue différents selon les secteurs d'activité.

Ce type d'initiative met en évidence l'engagement en faveur d'un dialogue constructif entre les institutions de régulation et les organisations privées. Il présente également l'avantage d'introduire le sujet d'une IA digne de confiance au sein des entreprises et de les y sensibiliser.

Citons ici quelques points remontés au HLEG en novembre 2019 :
Les tests du questionnaire HLEG sur des cas concrets ont montré trois lacunes principales qui le rendent insuffisamment opérationnel dans un contexte commercial : (a) le manque de clarté quant à l'utilisation du questionnaire, (b) le manque de considération des enjeux sectoriels de l'IA et (c) le manque d'actionnabilité du questionnaire.

(a) Selon les observations du groupe de travail IA Responsable, le questionnaire du HLEG aurait dû présenter des recommandations concrètes de mise en œuvre. À quel stade du cycle de vie d'une solution d'IA est-il pertinent de procéder à cet examen ? Quel service ou département de l'entreprise pour s'en charger ? En résumé : qui l'utiliserait et quand ?

(b) Par ailleurs, le questionnaire ne souligne pas que, les secteurs d'activité présentant chacun des spécificités, leurs enjeux éthiques diffèrent, et les critères d'évaluation pourraient par conséquent varier selon les contextes. Une première approche, selon le groupe de travail IA Responsable, consisterait à concentrer les efforts de gouvernance sur les secteurs déjà identifiés par l'initiative AI4EU afin d'élaborer des instruments de mesure distincts selon les particularités des entreprises concernées tant en termes d'activité que de types d'intelligence artificielle développés et d'usages prévus.

(c) Enfin, il semblerait judicieux de hiérarchiser les différents éléments d'une IA à examiner afin de les pondérer dans l'évaluation éthique globale du produit ou service. Selon le groupe de travail IA Responsable, identifier les points critiques soutiendrait les équipes opérationnelles de façon pratique et pertinente.

— ET DEMAIN ?

Début 2020, le groupe IA Responsable d'Impact AI a décidé d'amorcer un cycle de travail sur la gouvernance de l'intelligence artificielle. Celui-ci teste, à partir de trois ateliers, une approche bottom-up pragmatique qui souligne les enjeux de l'autorégulation des entreprises sur la problématique, à contre-pied de l'option d'évaluation top-down des risques de l'IA portée par l'AI HLEG.

Ce cycle de travail a permis aussi de présenter et de partager les convictions d'Impact AI sur la gouvernance, en particulier les bonnes et mauvaises pratiques, afin de soutenir les entreprises dans leur cheminement vers la mise en place d'une intelligence artificielle digne de confiance (voir p. 70).

Ce guide pratique s'inscrit dans cette démarche et publie le fruit du travail des trois ateliers réunis en 2020.

À RETENIR

Il existe des risques concrets et spécifiques aux solutions basées sur l'IA, pour la société comme pour les acteurs qui les produisent.

Afin de faire émerger une IA responsable, les organisations doivent mettre en place une gouvernance spécifique, adopter une démarche d'éthique *by design* et favoriser l'acculturation des utilisateurs.

Les autorités, notamment européennes, s'efforcent d'élaborer un cadre normatif propre à créer une IA digne de confiance sans entraver son développement.

Impact AI a testé en 2019 le guide d'évaluation éthique d'une IA proposé par les instances européennes et formulé trois propositions dans un « *Policy Paper* » :
— Préciser qui doit examiner une IA sous son angle éthique et à quel stade de son cycle de vie.
— Adapter les grilles d'évaluation aux spécificités des secteurs d'activité.
— Hiérarchiser les critères d'évaluation.

En 2020, Impact AI a développé une approche d'identification des besoins et des dispositifs actuels de gouvernance de l'IA au sein des entreprises françaises. Ce guide recense ce travail et présente des bonnes pratiques à considérer pour toute organisation.

2

CHAPITRE

Des principes généraux à la pratique

Quelles règles d'éthique pour soutenir une stratégie d'intelligence artificielle digne de confiance et comment les déterminer ? De quelle manière les articuler avec les risques que fait peser l'IA sur les organisations ? Et avec les recommandations des instances normatives ? Le groupe de travail IA Responsable a déterminé sept principes-clés passés au crible de l'expérience de ses membres. Le résultat de cette démarche empirique esquisse l'état de l'art de leurs connaissances et… des lacunes à combler.

DÉFINIR DES RÈGLES ÉTHIQUES, C'EST À L'ÉVIDENCE COMMENCER PAR SPÉCIFIER ET CIRCONSCRIRE LES RISQUES LIÉS À L'INTELLIGENCE ARTIFICIELLE. L'analyse s'effectue sous deux angles : la probabilité qu'un problème survienne d'une part, son degré de gravité s'il se produit d'autre part. Une fois achevée cette évaluation combinant ces deux facteurs, l'organisation peut s'attacher à concevoir des mécanismes de protection efficaces. Elle doit à ce stade tenir compte de son système préexistant de gestion des risques. En effet, il existe bien souvent des organes et procédures en la matière au sein des entreprises qu'il serait contreproductif d'empiler avec de nouveaux dispositifs. Il faut au contraire y intégrer les questions relatives à l'IA afin d'éviter les redondances et d'assurer une cohérence globale de la stratégie.

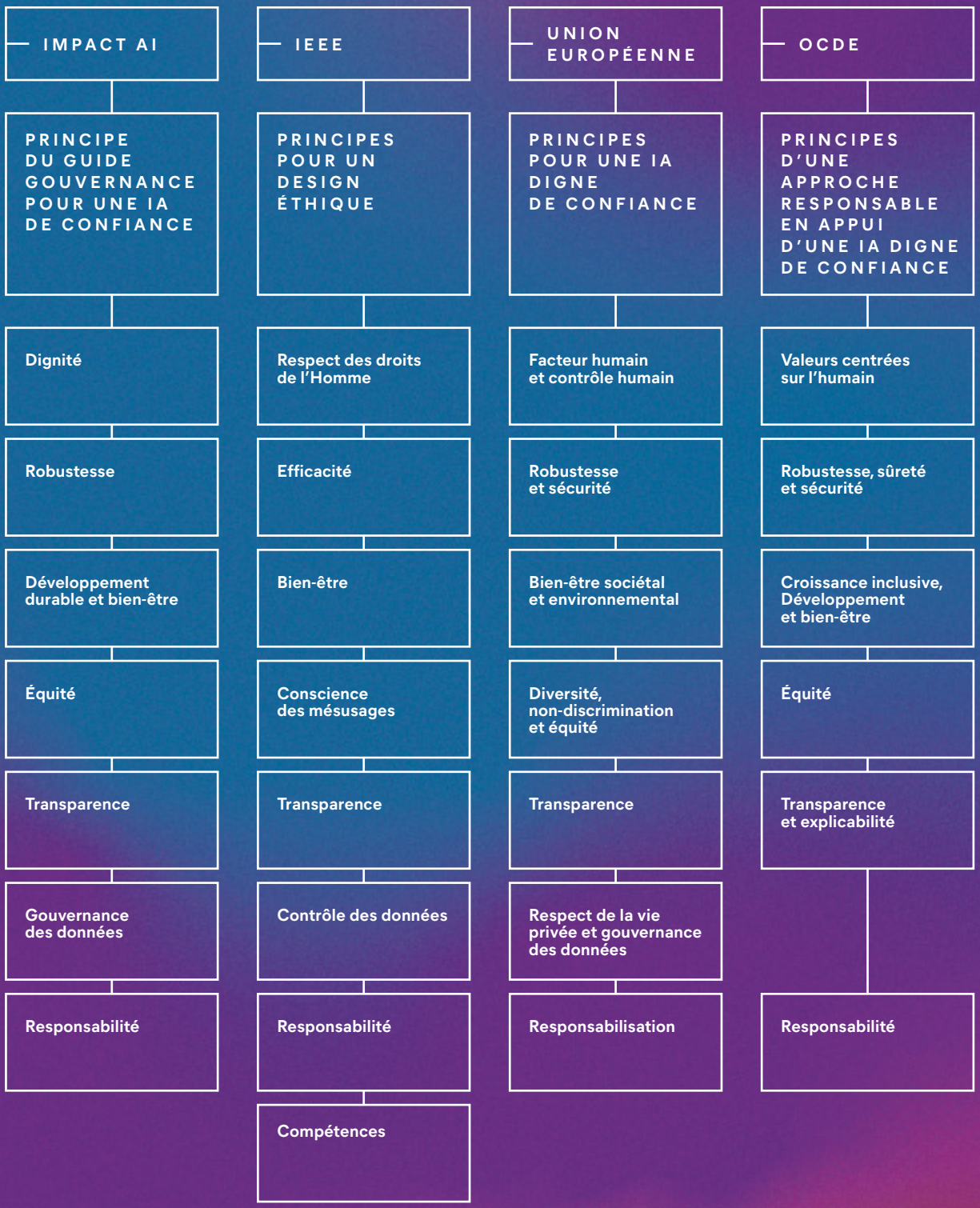
Enfin, la réflexion sur la mise en place d'une gouvernance de l'intelligence artificielle responsable associe trois méthodes :

<1>. **La première** consiste à se questionner « au bon moment » : il s'agit de déterminer quelles seront les questions pertinentes et nécessaires à se poser à tel ou tel stade du cycle de vie d'une application. Ainsi, il est possible de s'interroger à l'étape de la conception sur le caractère privé ou non des données à utiliser et sur la manière de garantir leur protection (anonymisation, chiffrage, agrégation) ;

<2>. **La deuxième** est l'automatisation de procédures qui techniquement empêchent le développement et le déploiement d'un outil d'intelligence artificielle s'il ne satisfaisait pas à des critères prédéterminés. Par exemple, il peut être décidé que certaines données ne sortiront d'une enceinte que sous une forme anonymisée, agrégée ou chiffrée ;

<3>. **En troisième** lieu, si l'organisation a établi des référentiels adaptés à chaque contexte, la certification et l'audit peuvent compléter ces dispositifs.

<fig.3.> Les principes du « Guide pratique pour une IA digne de confiance » s'inspirent de ceux établis par l'IEEE, l'Union européenne et l'OCDE.



Que la voie choisie combine les trois options ou qu'elle ne s'appuie que sur l'une d'elles, la formation des différentes parties prenantes demeure toujours une nécessité. Ces constatations posées, le groupe de travail s'est inspiré des principes éthiques édictés par l'IEEE, l'Union européenne et l'OCDE et les a synthétisés en sept mots-clés qui forgent l'ossature d'un cadre de gouvernance pour toute organisation : dignité, robustesse, développement durable et bien-être, équité, transparence, gouvernance des données et responsabilité.

S'accordant sur ces concepts, les membres ont examiné les conditions méthodologiques et techniques de leur mise en œuvre avant de passer en revue leurs pratiques respectives.

Dignité : une IA au service de l'humain

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : les systèmes d'intelligence artificielle devraient être les vecteurs de sociétés équitables en se mettant au service de l'humain et des droits fondamentaux, sans restreindre ou dévoyer l'autonomie des individus.

Imaginer un produit ou un service d'IA sur lequel l'Homme n'a pas le contrôle final tient de la dystopie. À l'évidence, toute solution partiellement ou totalement automatisée pose comme un impératif la maîtrise *in fine* par l'utilisateur. Cette exigence est de ce fait constitutive de l'expression du besoin et de la phase de conception de l'application.

— Retour d'expérience

Axionable, spécialiste du conseil en intelligence artificielle durable et responsable, a systématisé l'utilisation d'un « business canevas » qui met en regard les buts de l'application d'intelligence artificielle en projet et ses dimensions responsables et durables. Cet outil est coconstruit et validé par l'ensemble des parties prenantes du programme. Une modélisation simple de l'algorithme s'ensuit qui précise les données d'apprentissage, les résultats attendus et les contraintes à respecter, notamment en termes d'interaction avec l'utilisateur.

De manière similaire, les documents de spécification des solutions d'intelligence artificielle de **Schneider Electric** récapitulent l'objectif *business,* les besoins de formation de l'utilisateur, le niveau d'implication et de contrôle souhaitables pour celui-ci et les interactions nécessaires pour y parvenir (présentation d'options et/ou d'explications pour une décision humaine, alerte en présence de données atypiques, etc.). Placer l'utilisateur et son contrôle de l'application au cœur de la conception : tel est le paradigme. La fonction de la technologie est de l'assister tout en le laissant maître de l'examen de ses données, de leur catégorisation ou de leur sélection pour un éventuel réapprentissage par l'algorithme.

Deux difficultés de poids subsistent néanmoins pour parvenir à un usage d'une application parfaitement contrôlée par son bénéficiaire. D'abord, dans certains cas, il peut s'avérer ardu de concevoir des interfaces qui rendent compte avec clarté de la logique appliquée par le système. Et dans ces conditions, comment être certain qu'un utilisateur, même formé, prenne des décisions éclairées ?

Par ailleurs, un algorithme de machine learning a vocation à proposer des recommandations ou des choix qui conviennent à l'utilisateur ; cette pertinence éprouvée peut émousser sa vigilance et son esprit critique et renforcer des réflexes de passivité jusqu'à laisser un jour le système faire un choix discutable.

D'une manière générale, ces obligations autant que ces incertitudes soulignent la place centrale que tient la pédagogie pour bâtir une IA digne de confiance. Celle-ci se façonne dès le stade de la conception du système d'intelligence artificielle par la réflexion sur des interfaces claires qui garantissent l'autonomie de la décision de l'utilisateur et réduisent ses risques d'erreur. Elle se prolonge ensuite par l'élaboration de formations idoines.

Robustesse : une IA fiable dans la durée

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : une IA digne de confiance nécessite des algorithmes suffisamment sûrs et fiables pour gérer les erreurs ou les incohérences à toutes les étapes du cycle de vie d'un outil d'intelligence artificielle.

Un algorithme de machine learning trie, classe ou ordonne des informations selon certains principes afin de produire un résultat qui lui a été spécifié. Mais il lui arrive de commettre des erreurs que la robustesse technique et la sécurité visent à prévenir. L'objectif des concepteurs consiste à ce que, tout au long de son cycle de vie, le système se comporte de manière fiable, c'est-à-dire comme prévu, en atténuant les dommages involontaires et inattendus et en évitant ceux jugés inacceptables.

Pour satisfaire ce besoin, le créateur d'un algorithme d'apprentissage doit pouvoir répondre à deux questions. Comment éviter que l'application d'intelligence artificielle produise des erreurs d'une part ? D'autre part, comment, si celles-ci adviennent, limiter leurs conséquences, notamment lorsque la sécurité des personnes, des biens ou de l'environnement est en jeu ?

Les équipes de sécurité de DeepMind résument en trois critères une solution d'intelligence artificielle fiable et solide : spécification — que doit produire l'application ? —, robustesse — quelle conception pour éviter toute perturbation ? —, assurance — quels moyens de pilotage et de contrôle du système ?

— Retour d'expérience

Pour éviter, détecter et corriger les erreurs potentielles, Axionable découpe en trois temps son travail qui épouse ainsi le cycle de développement d'une solution d'intelligence artificielle.

En amont, lors de la conception, des méthodes et grilles de critères spécifient en détail le comportement attendu de l'algorithme afin de supprimer toute ambiguïté et d'anticiper les possibles dérives et les effets de bord[26].

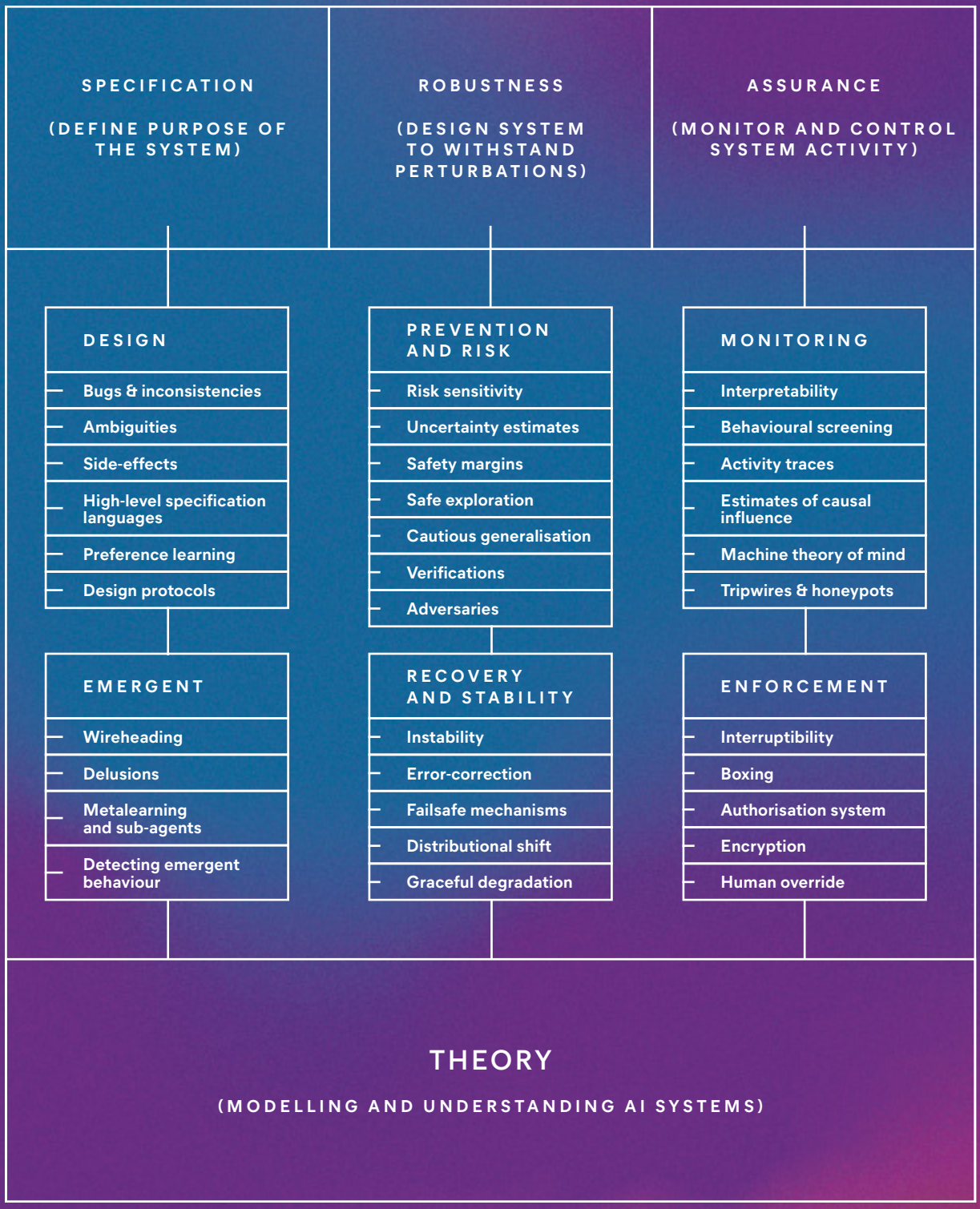
Avant la mise en production de l'application, des problématiques types, préalablement identifiées par les *data scientists* ou *data engineers,* sont testées sous la forme de scénarios. Ces essais génèrent des ensembles de données synthétiques qui, à leur tour, permettent de jouer des scénarios aléatoires et/ou incohérents.

Grâce à cette double approche, Axionable vérifie non seulement l'adéquation entre le comportement de l'algorithme et ses spécifications, mais encore détecte les cas extrêmes qui n'avaient pas été anticipés. C'est ainsi que l'on étudie par exemple la réaction du

25. Pedro A. Ortega, Vishal Maini, and the DeepMind safety team, « Building Safe Artificial Intelligence: Specification, Robustness, and Assurance », Medium, septembre 2018

26. Pour la définition des termes techniques, voir le glossaire p. 94

<fig.4.> Les trois piliers d'une IA solide et fiable selon DeepMind[25]



système lorsqu'il reçoit en entrée une donnée qui n'avait pas été prévue lors de la conception. Cette méthode par série de tests exige une gestion rigoureuse des différentes versions de l'application d'IA et des hyperparamètres associés. L'algorithme n'est mis en production que lorsque toutes les étapes ont été validées.

Après la mise en production, la surveillance du maintien de la performance de l'application d'intelligence artificielle au cours du temps est capitale car celle-ci a généralement tendance à décroître. Pour éviter ce problème, les équipes s'appuient sur un système d'alertes lancées dès que des seuils prédéfinis sont franchis et sur trois méthodes qui détectent puis atténuent ou corrigent les erreurs :

<1>. D'abord le *ground truth* ou vérité de terrain en français. Cet indicateur compare les performances des algorithmes telles qu'elles avaient été prévues au moment de la conception, avec celles qui sont effectives lorsqu'elles reposent sur des données réelles.

<2>. Ensuite, le *data drift* (glissement de données) pose que si les données fournies à l'algorithme évoluent au cours du temps, celui-ci n'y est a priori pas adapté et l'exactitude de ses prédictions s'en trouve détériorée. Par exemple, une solution de machine learning de domotique qui régule le chauffage d'un appartement peut avoir été programmée pour analyser des données de températures extérieures variant selon des moyennes saisonnières. Si des températures extrêmes se présentent, le réglage du chauffage risque d'être sous-optimal car le système n'a pas été conçu ou entraîné pour ce nouveau type de données.

<3>. Enfin le *concept drift* apparaît lorsque l'objectif initial fixé à l'application se trouve modifié. Là encore, la fiabilité du résultat est altérée. Prenons un algorithme dont la fonction originelle consiste à calculer le prix de téléphones mobiles d'occasion en se fondant sur la marque, le modèle et la date initiale de mise sur le marché. Si un critère supplémentaire d'évaluation apparaît tel que le prix de vente chez les concurrents, les résultats fournis par l'application risquent de se révéler inexacts.

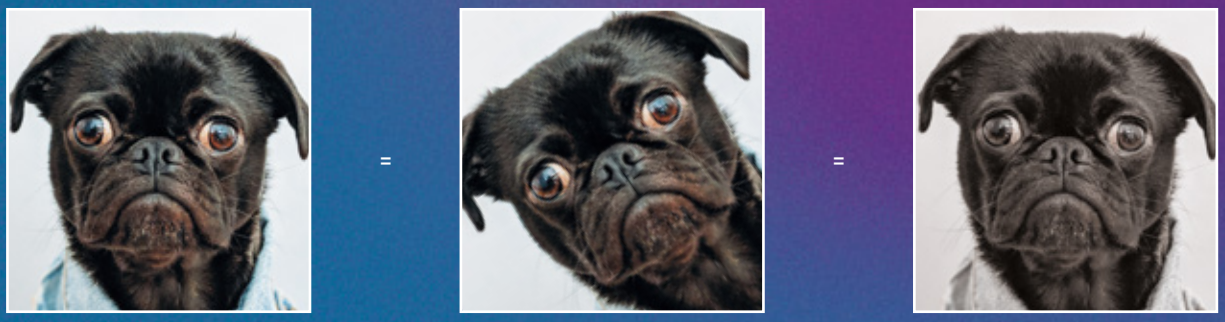
De son côté, Schneider Electric met en œuvre une méthodologie préventive graduelle selon le domaine d'application, les conséquences possibles d'une erreur, les propriétés du modèle d'IA et le rôle de l'utilisateur.

La méthode de mesure de la fiabilité d'un modèle appris constitue le socle de la démarche : toute solution d'IA est documentée sur ce point. L'évaluation est réalisée systématiquement pour chaque modèle appris et par des tests unitaires et fonctionnels en cas d'évolution de l'application. À cette procédure s'ajoutent parfois un ensemble d'alarmes, plus ou moins complexes, qui détectent les situations où l'algorithme est confronté à des combinaisons de données qu'il connaît mal, sous-représentées durant la phase d'apprentissage.

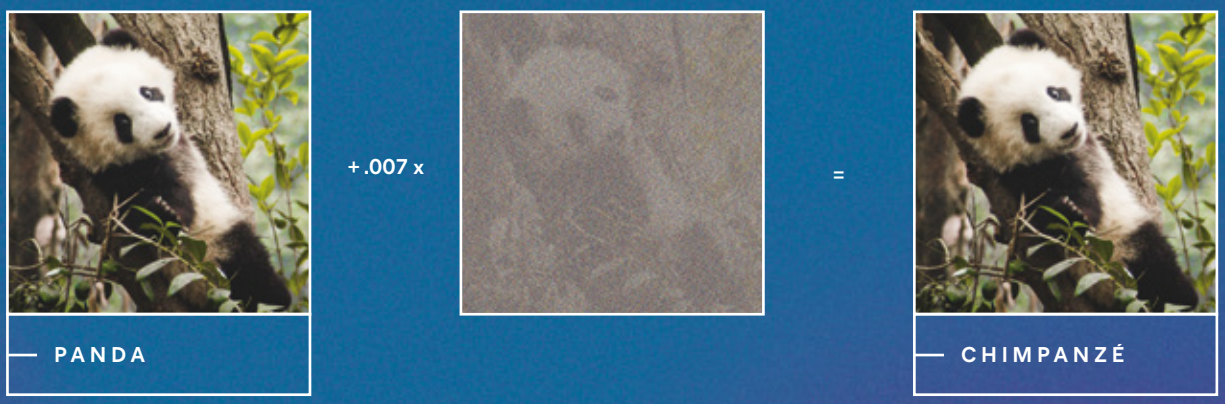
Dans de nombreux cas, une solution de sécurité peut cohabiter avec l'application d'intelligence artificielle et s'activer en cas de glissement des données. Pour ne donner qu'un exemple, on peut très bien utiliser une application d'intelligence artificielle pour détecter dès que possible l'apparition d'un défaut dans le comportement d'un appareil et maintenir une procédure d'arrêt automatique de l'appareil si la température mesurée dépasse un seuil donné, inhabituel et potentiellement dangereux.

27. Ian J. Goodfellow, Jonathon Shlens, et Christian Szegedy, « Explaining and Harnessing Adversarial Examples », *arXiv:1412.6572 [cs, stat]*, mars 2015

<fig.5.> L'algorithme a appris à partir de photos de chien nombreuses et diverses. Une fois en production, il est confronté à une image légèrement modifiée et reconnaît néanmoins le chien : il est réputé robuste.



<fig.6.> Des images légèrement modifiées par rapport au modèle d'apprentissage produisent des incohérences dans les résultats de l'algorithme qui les classe dans des catégories erronées. L'algorithme n'est pas robuste[27].



<fig.7.> L'algorithme prédit correctement que l'image originale (à gauche) représente un panda avec un taux de confiance de 57 %. Mais lorsque que l'on ajoute un « bruit adverse » (adversarial noise), l'algorithme prédit cette fois à tort qu'il s'agit d'un chimpanzé, avec un taux de confiance nettement plus élevé de 99 %.

Les technologies de machine learning apprennent à effectuer une tâche à partir d'exemples réels ; la stabilité de leurs résultats lorsque des données sont légèrement modifiées témoigne de leur robustesse. Pour parvenir à cet objectif, l'apprentissage doit se baser sur la plus grande diversité possible d'exemples. À défaut, l'algorithme se fondera sur des raccourcis pour établir des prédictions qui de ce fait se révèleront erronées. Par exemple, s'il a été entrainé uniquement avec des images de chats blancs, il ne saura pas reconnaitre un chat noir. Ce qu'il aura appris à reconnaitre ne sera pas généralisable. La robustesse d'un outil d'IA est donc indissociable de sa capacité de généralisation. Dans de nombreux domaines, un algorithme est dit robuste lorsque sa prédiction reste correcte en dépit de la modification d'un exemple.

Or, des travaux de recherche ont démontré que les prédictions de modèles courants de machine learning ne sont pas toujours stables dès lors qu'un exemple est modifié. Ainsi, flouter une image, remplacer un mot par un synonyme ou ajouter un léger fond sonore peut significativement altérer la classification effectuée par l'algorithme de l'image, de la phrase ou du son. Il est même possible de tromper l'outil en ajoutant à une image une perturbation spécifique, souvent indécelable. Les résultats produits sont aberrants en raison de ces adversarial attacks.

Deux types de techniques peuvent être adoptés pour remédier à ce problème. D'abord, des travaux scientifiques affirment que l'introduction d'images perturbées et modifiées durant le processus d'apprentissage contribue à limiter les risques d'erreurs. Ensuite, il est possible, toujours à ce stade du cycle de vie du modèle d'IA, de détecter le moment où celui-ci se fourvoie de telle sorte qu'un humain puisse prendre la relève et livrer la réponse exacte. C'est sur cette méthode que se fonde le pourcentage de confiance accordé aux prédictions d'un algorithme. Cela n'exclut pas toutefois les anomalies puisque des *adversarial attacks* postérieures à la phase d'apprentissage peuvent encore brouiller le résultat. Il est donc indispensable de mesurer la fiabilité du taux de confiance avant de déclarer qu'un algorithme fournit des réponses pertinentes et qu'il est donc robuste.

Il existe d'autres difficultés liées à la robustesse des algorithmes de machine learning telle que la fiabilité des prédictions.

Si d'autres systèmes informatiques sont également concernés par ce type de problématique, l'intelligence artificielle a ceci de spécifique qu'elle apprend. Et pour cela elle utilise des données du passé dans le but de prédire le futur ou de fournir des résultats qui seront utilisés dans un autre contexte.

— Retour d'expérience

Pour traiter de la question de la robustesse, AXA investit dans la recherche et ses équipes de R&D[28] publient régulièrement des articles scientifiques évalués par des pairs. L'un d'eux a révélé qu'il est possible, dans certains cas, de calculer une mesure de confiance afin de détecter de manière fiable les erreurs, les images hors-distributions et les *adversarial attacks* dans un algorithme de machine learning. Un autre article détaille comment générer des *adversarial attacks* pour des données tabulaires, c'est-à-dire représentables sous forme de tableau, qui sont les plus utilisées chez AXA.

28. Pour en savoir plus sur les recherches d'AXA sur la robustesse

La robustesse d'un outil d'IA est indissociable de sa capacité de généralisation.

DANS DE NOMBREUX DOMAINES, UN ALGORITHME EST DIT ROBUSTE LORSQUE SA PRÉDICTION RESTE CORRECTE EN DÉPIT DE LA MODIFICATION D'UN EXEMPLE.

Gouvernance des données : une IA respectueuse de la vie privée

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : ensemble des mesures qui visent à s'assurer de la qualité et de l'intégrité des données utilisées ainsi que de leur pertinence dans le domaine d'application considéré. Il s'agit aussi d'établir des protocoles d'accès à ces données et des règles pour les traiter qui garantissent le respect de la vie privée.

— Retour d'expérience

Deloitte a mis en place une méthodologie conforme au Règlement général de la protection des données (RGPD) qui préserve la confidentialité des données de ses clients tout au long du cycle de vie d'une application d'intelligence artificielle.

Elle ordonnance la collecte des données stratégiques et assure leur intégrité dans le processus d'apprentissage. Toutes les procédures et tous les ensembles de données sont testés et documentés, à toutes les étapes du cycle de vie du système d'IA : conception, collecte, préparation, traitement, essais, exploitation et pilotage. Des protocoles définissent également les habilitations d'accès aux données et les documents pour les parties prenantes. Ces dispositions s'appliquent aux outils développés en interne mais aussi en externe et maintiennent ainsi la confiance des utilisateurs.

Prise individuellement, une donnée ne présente aucun intérêt ; c'est l'agrégation puis le partage qui créent la valeur. Pour autant, la protection de la vie privée et de la confidentialité des affaires constitue une norme nécessaire. Pour concilier ces deux impératifs légitimes, la politique d'accès aux données de Microsoft est sous-tendue par le principe « aussi ouvert que possible, aussi fermé que nécessaire ». Ceci se matérialise dans des outils qui facilitent la tâche des organisations et des individus désireux de partager des données. Parmi ceux-ci, quatre types de contrats qui peuvent sécuriser les acteurs et leurs opérations.

<1>. *L'Open Use of Data Agreement* (O-UDA) concerne un individu ou une organisation détenteur de données dont il autorise l'exploitation illimitée. Ce contrat *one-to-many* porte sur des informations publiques. Par exemple une autorité portuaire qui collecte des indications météorologiques — direction et vitesse du vent, températures, précipitations… — pourrait rendre ces renseignements accessibles au public grâce à l'O-UDA. Ces données, ni sensibles ni privées, pourraient être ainsi librement utilisées.

<2>. *Le Computational Use of Data Agreement* (C-UDA) s'applique lorsqu'il est question d'exploiter des données collectées légalement par une organisation à partir de sources ouvertes au public. Cet accord permet d'ouvrir l'accès à ces informations mais uniquement à des fins de calculs statistiques, comme le fait un algorithme d'apprentissage. Il s'agit d'un accord *one-to-many* utilisable à la condition que les données ne soient pas confidentielles. Par exemple, une organisation peut mettre à disposition de tous sa banque d'images qu'elle a constituée légalement à partir de sources accessibles au public mais qui contient potentiellement des photos protégées par le droit d'auteur. Le C-UDA

l'autorise à rendre ces données accessibles au public pour une utilisation informatique uniquement, par souci de cohérence avec les lois sur le droit d'auteur.

<3>. *Le Data Use Agreement for Open AI* (DUA-OAI), ou accord d'exploitation de données pour le développement d'un modèle d'IA ouvert, est une troisième possibilité. Il prévoit les conditions de partage de données ouvertes au public entre deux organisations dont la seconde souhaite développer une application d'IA en *open source*[29]. Ce contrat *one-to-one* s'applique à des informations confidentielles ou touchant à la vie privée. Par exemple, une entreprise qui disposerait d'une base de données recueillies dans ses usines pourrait vouloir la proposer à autre société du même secteur d'activité dont le but serait de mettre au point un algorithme d'intelligence artificielle qui propose des améliorations des dispositifs de sécurité dans des sites de production. La base de données est susceptible de contenir des informations sensibles. En s'appuyant sur le DUA-OAI, le partage entre les deux structures est autorisé s'il donne lieu à la conception d'un modèle d'IA qui sera lui-même en libre accès.

<4>. L'accord d'utilisation des données pour les données communes *(Data Use Agreement for Data Commons* [DUA-DC]) est multipartite ou *many-to-many* en anglais. Il s'applique lorsque des détenteurs d'ensembles de données sur une thématique particulière souhaitent les mettre en commun pour les utiliser. Ce contrat prévoit que les parties constitueront la nouvelle base de données grâce à des interfaces de programmation (API) spécifiées et y auront accès par d'autres, elles-mêmes déterminées. Par exemple, des communes limitrophes souhaitent partager leurs informations sur leurs feux tricolores et accéder à cette base de données afin de réguler le trafic routier de toute la région. Les API garantissent un format de données exploitable par les différents systèmes de contrôle des feux de signalisation. Les parties fondent ce partage de données sur un contrat de type DUA-DC.

Transparence : une IA interprétable et explicable

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : les algorithmes d'apprentissage sont dits transparents lorsque le fonctionnement interne des modèles d'apprentissage sont rendus publics ; ils répondent à l'exigence d'explicabilité quand ces mêmes éléments restent intelligibles, c'est-à-dire qu'il est possible de comprendre à quelles règles ils obéissent.

Quelle que soit la qualité de leurs prédictions, les systèmes de machine learning peuvent engendrer des questions sur leur interprétabilité, au point de créer un rejet chez des utilisateurs, grand public ou entreprises. Ainsi, si un algorithme d'apprentissage du risque de crédit refuse à quelqu'un une demande de prêt et que les conseillers bancaires peinent à expliquer et justifier cette décision cela provoque une incompréhension justifiée de la part de la personne. L'algorithme d'apprentissage pèche par manque de transparence et d'explicabilité. Il s'agit donc, en s'appuyant sur un principe de transparence, de développer des applications d'intelligence artificielle qui permettent de comprendre les décisions ou prédictions du modèle d'IA.

29. Une application en open source est un programme informatique dont le code source est distribué sous une licence permettant à quiconque de lire, modifier ou redistribuer ce logiciel.

Les systèmes de machine learning peuvent engender des questions sur leur interprétabilité

AU POINT DE CRÉER UN REJET CHEZ DES UTILISATEURS, GRAND PUBLIC OU ENTREPRISES.

La transparence se joue d'abord vis-à-vis de l'utilisateur : il doit savoir qu'il interagit avec un outil d'apprentissage automatique fondé sur des données qu'il est souhaitable de caractériser, ces informations pouvant figurer de façon explicite dans la documentation qui lui est fournie ou dans la formation qu'on lui propose. Néanmoins, cela reste parfois insuffisant pour répondre aux questions de l'interprétabilité et de l'explicabilité. Autrement dit comment faire comprendre à l'utilisateur le fonctionnement du modèle et la manière dont il parvient à un résultat d'une part, et de quelle manière l'expliquer d'autre part.

— Retour d'expérience

Schneider Electric traite dès l'origine d'un projet d'intelligence artificielle les besoins de traçabilité et d'interprétabilité, la disparité des connaissances des utilisateurs et la nécessaire adéquation des outils aux compétences de chacun. Dans certains cas, une formation peut s'avérer nécessaire.

À performances égales, on peut préférer l'utilisation d'un modèle naturellement « simple », tel un arbre de régression. La capacité d'examen ou de traçage de l'algorithme de machine learning va néanmoins dépendre de son ampleur. Des systèmes plus complexes, comme ceux de *deep learning*, aujourd'hui les plus performants sur des données non structurées, sont très difficilement explicables.

Les concepteurs se trouvent confrontés à une double contrainte : assurer l'explicabilité de l'algorithme sans compromettre sa performance. Si les outils sont à l'origine conçus pour établir un arbitrage le plus équilibré possible entre ces deux dimensions, il apparaît toutefois important de réfléchir à un niveau de compromis selon la finalité de chacun d'eux.

Si l'on veut à la fois que le modèle soit efficace et performant il faut alors exploiter des méthodes d'interprétabilité plus sophistiquées.

— Retour d'expérience

L'approche de transparence de Deloitte intègre l'identification des besoins de compréhension de la part des différentes parties prenantes d'un projet d'IA ainsi que des méthodes pour répondre à ces besoins sur l'ensemble du cycle de vie des données et du système d'intelligence artificielle[30]. Ces procédures s'enrichissent de protocoles d'identification des biais discriminants (voir p. 60).

De nombreuses questions fondamentales trouvent réponse à l'aide de ces procédures :

<1>. Quelle est la fonction remplie par l'application ? Quelles sont ses retombées potentielles sur le contexte donné ? Quels sont les processus métiers impactés ?

<2>. Quelles sont les parties prenantes qui requièrent l'explicabilité (métiers, régulateurs, management, etc.) ?

<3>. Faut-il expliquer comment fonctionne l'application ou pourquoi elle prend telle décision ? Les deux ? Avec quel niveau de détail ?

<4>. Quels sont les besoins prioritaires et comment les satisfaire au mieux ? L'explication doit-elle être guidée par des politiques internes ou des cadres réglementaires généraux ?

30. World Economic Form and Deloitte (2020). Navigating Uncharted Waters : A Roadmap to Responsible Innovation with AI in Financial Services.

Grâce à cette démarche, managers, auditeurs ou d*ata scientists* sont sensibilisés et disposent de tous les éléments nécessaires à la prise de décisions éclairées. Ils peuvent en effet vérifier l'influence de chaque variable et améliorer la performance du modèle d'IA ; ils en analysent aussi le comportement localement (expliquer une prédiction spécifique grâce à des techniques comme LIME et SHAP par exemple) et globalement (expliquer la logique sous-jacente du système).
MAIF pour sa part inscrit la transparence au cœur de sa relation avec ses sociétaires et la communauté de développeurs. C'est pourquoi elle propose plusieurs outils d'intelligence artificielle en open source afin que chacun puisse les évaluer, s'en emparer et contribuer à leur amélioration. La mutuelle d'assurance a ainsi rendu public le code de sa solution de traitement des mails, Mélusine. Dans la continuité de cette action et sur la base des recommandations de l'Union européenne pour une IA digne de confiance, ses équipes ont établi un questionnaire pour identifier d'éventuels risques ou points faibles de ses modèles d'intelligence artificielle en termes d'équité, d'impact sur la société… Ces initiatives renforcent ses engagements en incluant les dimensions d'éthique et de transparence dans ses projets.

AXA s'est associée à l'OCDE pour expliquer les prédictions des algorithmes sur les causes des crises économiques en se basant sur le principe de la transparence. La compagnie soutient également les travaux d'une thèse à Sorbonne Université dont l'objet est d'étudier la manière optimale de faire comprendre aux néophytes les résultats d'une IA en se fondant sur les méthodes actuelles d'interprétabilité.

Équité : une IA qui traite chaque personne de manière juste

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : alors que l'on attend des résultats d'un algorithme de machine learning une neutralité parfaite, il arrive que ceux-ci soient biaisés. Un biais est un préjugé en faveur ou en défaveur d'une chose, d'une personne ou d'un groupe par rapport à un autre et qui est considéré comme injuste. L'IA équitable, débarrassée des biais, peut contribuer à réduire les inégalités et à bâtir une société plus inclusive.

Le machine learning est utilisé pour prendre des décisions dont les conséquences sont déterminantes pour les individus : acceptation ou refus d'une demande de prêt, publicité personnalisée, évaluation des CV… Le système doit donc impérativement accomplir ces tâches de manière équitable, c'est-à-dire sans discrimination en raison du genre, de l'origine réelle ou supposée, de la religion, l'orientation sexuelle etc. Or, une étude de Télécom ParisTech[31], par exemple, a démontré l'importance des biais dans les résultats des modèles d'apprentissage, que ceux-ci soient d'origine cognitive, statistique ou économique.

La problématique de l'équité de l'intelligence artificielle s'illustre à travers la polémique sur COMPAS[32] *(Correctional Offender Management Profiling for Alternative Sanctions),* algorithme de prédiction utilisé par le système judiciaire américain pour évaluer la probabilité de récidive d'un condamné sur la base d'une note fixée à partir d'informations personnelles ; plus le score est élevé, plus fort est le risque de récidive.

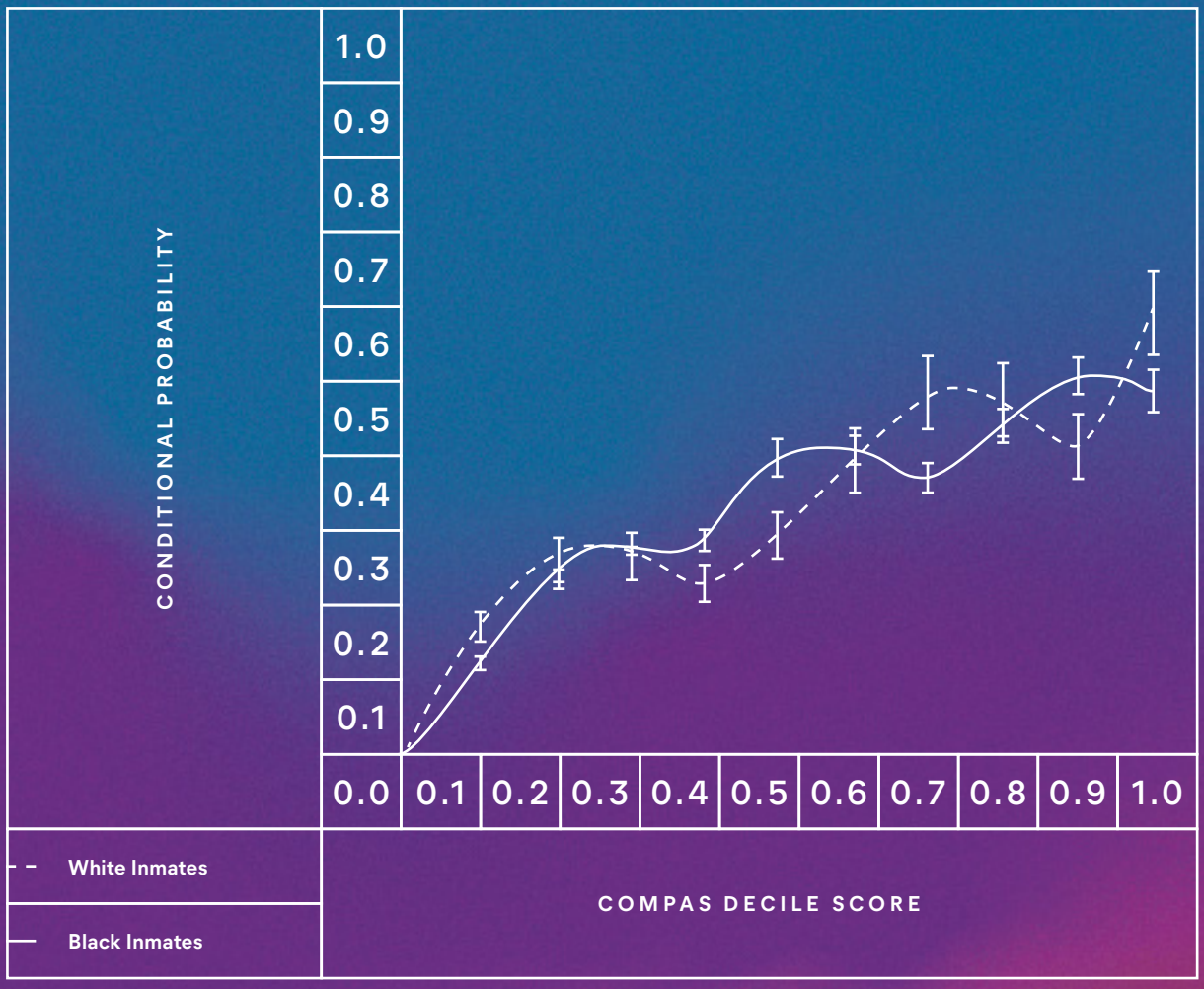
31. Patrice Bertail et al., « Algorithmes : biais, discrimination et équité », 2019

32. « L'efficacité d'un logiciel censé prédire la récidive à nouveau critiquée », Le Monde.fr, janvier 2018

<fig.8.> Les résultats par groupe ethnique de COMPAS.

PERSONNES DE COULEUR		
PERSONNES BLANCHES		
CONSIDÉRÉ COMME À HAUT RISQUE SANS AVOIR RÉCIDIVÉ	23,5 %	44,9 %
CONSIDÉRÉ COMME À FAIBLE RISQUE MAIS EN FAIT RÉCIDIVISTE	47,7 %	28,0 %

<fig.9.> Selon COMPAS, en fonction des scores obtenus par les personnes, les risques de récidives sont comparables d'un groupe ethnique à un autre.



En mai 2016, le site américain ProPublica publie un article intitulé *Machine bias* qui affirme que COMPAS est en réalité biaisé et défavorise les personnes de couleur. Lorsque l'on étudie les résultats de l'algorithme prédictif, disent les journalistes, les Afro-Américains sont plus souvent considérés comme étant hautement susceptibles de récidiver sans que cela soit effectivement observé ultérieurement : la probabilité est évaluée à deux fois plus que pour les personnes blanches, qui, à l'inverse, sont estimées moins risquées qu'elles ne devraient l'être. Northpointe, qui a développé et distribue l'algorithme, conteste les méthodes de ProPublica et réplique que son algorithme est équitable en avançant deux arguments principaux.

D'abord, COMPAS affiche la même précision pour les personnes noires que pour les blanches : les prédictions sont de 69 % quelle que soit la couleur de peau des individus.

Par ailleurs, les résultats de COMPAS sont « calibrés », ce qui signifie que pour tout score donné, la probabilité réelle que le prisonnier récidive est approximativement la même pour les deux groupes. Le graphique p. 57 montre la relation entre les scores COMPAS en abscisse et les réelles récidives en ordonnée. Il apparaît que, sur une échelle de 1 à 10, les détenus noirs et blancs dont la note COMPAS est de 6 présentent une probabilité de récidive entre 40 et 50 %, alors que si elle baisse à 3, le risque s'élève à 30 %, quel que soit le groupe.

Cette polémique illustre toute la difficulté à déterminer précisément ce que sont des modèles d'intelligence artificielle non-biaisés par des stéréotypes et donc non-discriminatoires. À l'occasion des évaluations régulières dont les algorithmes d'apprentissage font l'objet, il importe de s'assurer qu'ils ne reproduisent pas ni n'aggravent les discriminations sociales préexistantes.

Mais comment évaluer l'équité d'un modèle de machine learning ? Pour répondre à cette question de fond, encore faut-il avoir préalablement déterminé avec précision ce qu'est l'équité. Lorsque l'on songe aux débats des philosophes à ce sujet, on mesure la difficulté de la tâche.

À supposer qu'un consensus ait été trouvé sur cette définition, comment la traduire concrètement, c'est-à-dire techniquement, dans un algorithme d'apprentissage ? Jusqu'ici, un système d'intelligence artificielle était apprécié en fonction de la qualité de sa précision. Il s'agit donc de concevoir des modèles qui perpétuent cette exigence tout en y ajoutant celle de l'équité, nouvel impératif qu'il faut pouvoir mesurer pour le faire respecter.

— Retour d'expérience

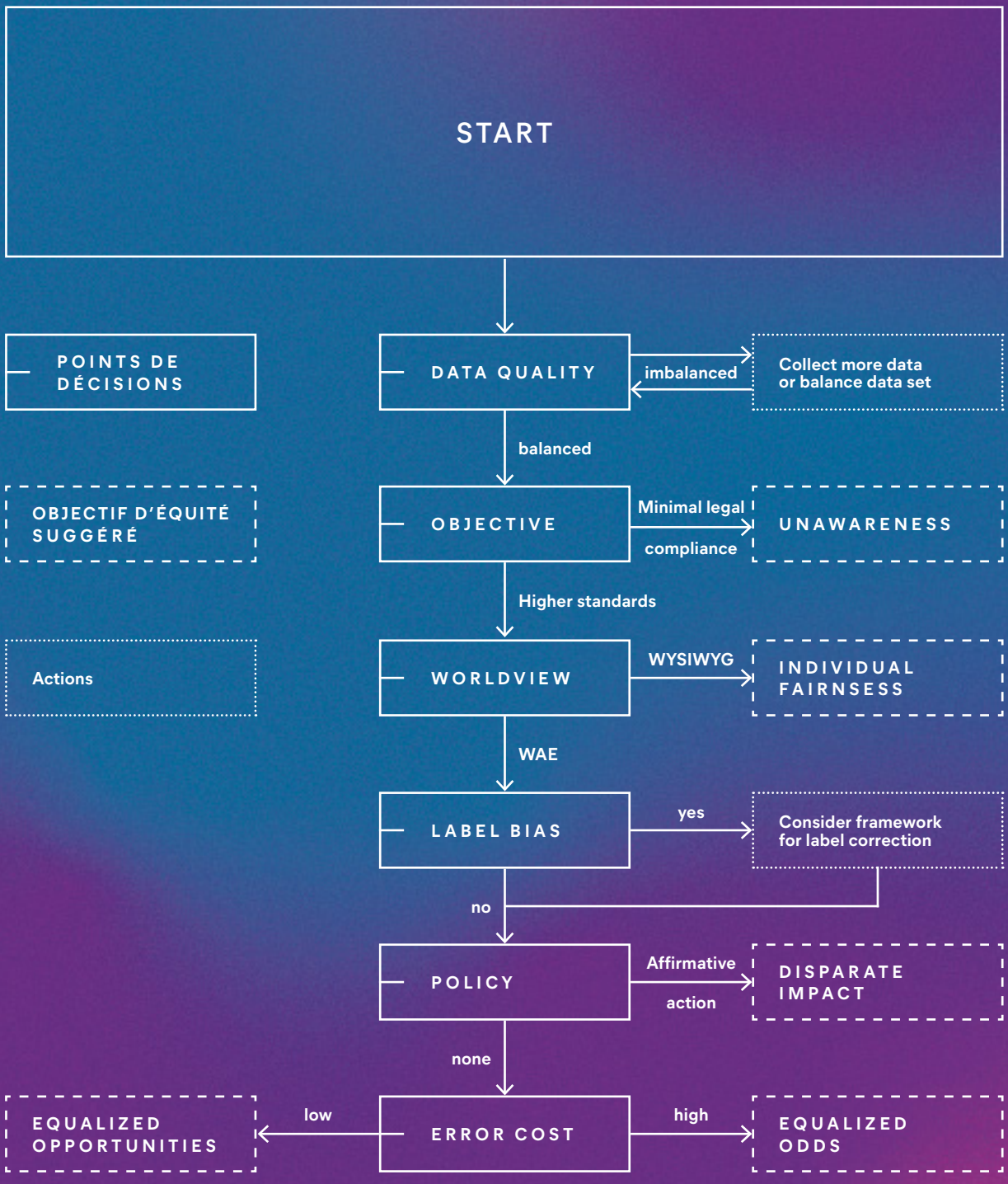
AXA considère les technologies de machine learning comme de formidables instruments de progrès de l'humanité dans toutes sortes de domaines. Mais elles ne sont pas dénuées de risques de discrimination car l'apprentissage automatique repose sur des ensembles de données, vastes, complexes et constitués dans le passé. Si historiquement ces informations étaient porteuses de préjugés, le modèle d'intelligence artificielle est susceptible de les reproduire, devenant de ce fait discriminatoire. L'entreprise mène de nombreuses recherches, publiées dans des revues scientifiques, pour développer des systèmes automatisés qui allient performance et impartialité des algorithmes d'apprentissage[33].

Afin d'approfondir la problématique de l'équité des solutions d'IA, l'équipe recherche et développement d'AXA Group Operation a développé le modèle ci-après (fig. 10) en proposant une boîte à outils pour structurer le paysage existant des définitions de l'équité les plus couramment utilisées. Cette proposition a été approfondie dans un article en libre accès[34].

33. Pour en savoir plus sur les recherches d'AXA dans ce domaine

34. Boris Ruf, Chaouki Boutharouite, et Marcin Detyniecki, « Getting Fairness Right: Towards a Toolbox for Practitioners », arXiv:2003.06920 [cs], mars 2020 - Pour consulter l'article

<fig.10.>[34] Proposition d'une boîte à outils pour structurer le paysage existant des définitions de l'équité les plus couramment utilisées.



Orange pour sa part s'est engagée sur la voie de l'équité en signant une charte sur l'IA inclusive avec Arborus en avril 2020[35]. Il s'agit d'un document de référence pour les entreprises de la Tech et pour celles qui sont utilisatrices de solutions d'intelligence artificielle. Les organisations s'engagent ainsi à respecter la diversité sur l'ensemble de la chaîne de la valeur de la donnée ainsi qu'à identifier et maîtriser les éventuels biais discriminatoires. Pour Orange, cela constitue une première étape avant l'obtention du GEEIS-IA, premier label international associant les enjeux de l'IA, la lutte contre les biais et stéréotypes en promouvant la diversité et l'usage responsable.

De nombreuses entités et fonctions sont ainsi mises à contribution dans ce cadre : la Diversité & Inclusion Groupe pour la politique d'égalité professionnelle entre les femmes et les hommes, la fonction RH pour qu'elle introduise cette problématique dans les processus de recrutement, la communication interne et externe pour diffuser l'information, les équipes formation pour qu'elles intègrent les risques de discrimination dans leurs corpus de programmes sur l'IA, les salariés sensibilisés aux stéréotypes et à leurs conséquences, les équipes chargées des relations sociales, celles des achats pour qu'elles prennent en compte cette question dans leurs échanges avec les fournisseurs et enfin les équipes-projets conceptrices ou utilisatrices d'applications d'intelligence artificielle. L'enjeu de sensibilisation de tous à l'IA responsable, quel que soit son métier au sein de l'entreprise est donc clé, et constitue un pilier par exemple des programmes d'acculturation à l'IA chez Orange.

Par ailleurs, le faible taux de représentation des femmes dans les écoles ou les fonctions d'ingénieur compliquent les efforts pour un recrutement paritaire. La conception d'un système d'IA équitable impose aussi d'effectuer des choix en fonction de contextes donnés : il importe que chaque équipe puisse arbitrer en connaissance de cause afin d'assumer ses décisions ensuite. Enfin, il n'est d'équité possible sans garantie sur la qualité des données.

Pour gérer le risque de biais, une des pistes est de s'appuyer sur un outil de visualisation comme IBM Fairness 360 qui propose de sélectionner des seuils à atteindre et des métriques. Il apparaît toutefois nécessaire de définir un processus qui aide à réaliser les bons choix, c'est ce qu'étudie Orange.

Le Cercle InterElles développe également une méthodologie d'aide aux entreprises engagées dans une démarche de non-discrimination de genre. Cette association professionnelle qui réunit les réseaux mixité de 15 grandes entreprises œuvre depuis 18 ans pour la féminisation et l'égalité professionnelle dans les secteurs scientifiques et technologiques. Plus de la moitié des entreprises partenaires sont à la fois utilisatrices et productrices d'outils d'intelligence artificielle.

Le groupe Femmes et IA du Cercle souhaite sensibiliser les organisations sur les discriminations de genre, les engager sur la voie d'une intelligence artificielle égalitaire et les rendre exemplaires en la matière. Or, s'attaquer à ce type de discrimination est techniquement relativement simple puisque les données de genre sont accessibles et de qualité dans toute organisation. En contribuant à l'écosystème français soucieux d'intégrer la dimension éthique à l'intelligence artificielle, le groupe Femmes et IA veut s'assurer que la contribution des femmes au développement de l'intelligence artificielle est considérée comme un sujet majeur par les organisations et l'État.

35. Arborus Orange, « Charte internationale pour une I.A. inclusive », avril 2020

<fig.11.> Les sept engagements d'Orange dans le cadre de la charte signée avec Arborus.

Opérationnellement, le Cercle InterElles propose aux entreprises un plan d'action en quatre étapes : s'engager par la signature d'une charte, s'évaluer grâce à une grille de critères, agir en puisant dans une boîte à outils enrichie régulièrement et enfin être visible et attractive via l'obtention d'un label.

Développement durable et bien-être : une IA qui contribue à résoudre des problèmes universels

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : les systèmes d'intelligence artificielle devraient être utilisés pour soutenir des évolutions sociales positives et renforcer la durabilité et la responsabilité écologique.

— Retour d'expérience

Axionable regrette que la multitude des référentiels — ceux définis en interne par les entreprises, celui de l'Union européenne, les 17 objectifs de développement durable de l'ONU —, complique la mesure de l'impact social et environnemental des algorithmes d'apprentissage. Par ailleurs, ces normes répondent à des définitions variables et leurs périmètres d'application diffèrent. Par exemple, le mot « absentéisme » n'a pas la même acception dans tous les pays. La classification des activités dites à hauts risques comme la reconnaissance faciale ou les applications de santé, actuellement en discussion à la Commission européenne, posera les premiers jalons d'un cadrage bienvenu.

Responsabilité : une IA capable de rendre des comptes

Définition connexe à celles de l'IEEE, de la Commission européenne et de l'OCDE : le développement de l'intelligence artificielle doit s'accompagner de garanties suffisantes pour minimiser le risque de dommages qu'elle peut causer. Il convient de mettre en place des mécanismes à cette fin et de les soumettre à l'obligation de rendre des comptes.

— Retour d'expérience

La méthodologie mise au point par Deloitte vise à identifier les risques d'un modèle d'IA, à réduire le plus possible ses effets délétères et à anticiper les besoins de conformité à venir définis par les autorités comme la CNIL, l'ACPR (Autorité de contrôle prudentiel et de résolution), l'Union européenne… Ces procédures, qui s'adaptent à toutes les étapes du cycle de vie des algorithmes de machine learning, ont permis de sensibiliser avec succès les clients et de les responsabiliser sur leur rôle au sein de la gouvernance de l'IA.

À RETENIR

- Le groupe de travail IA Responsable a identifié les questions à se poser avant toute initiative visant la mise en place d'une gouvernance.
- En corrélation avec les principes définis par l'Union européenne, l'OCDE et l'IEEE, il a déterminé sept piliers sur lesquels bâtir une intelligence artificielle digne de confiance : dignité, robustesse, gouvernance des données, transparence, équité, développement durable et bien-être, responsabilité.
- Les membres ont témoigné de leur engagement sur ces sujets en partageant leurs expériences, que celles-ci soient d'ordre organisationnel, juridique ou technique.
- Les solutions apportées pour respecter les principes éthiques soulèvent parfois de nouvelles questions, notamment scientifiques, sur lesquelles se mobilisent les structures de recherche des entreprises membres du groupe.
- Les piliers « développement durable et bien-être » et « responsabilité » méritent d'être creusés davantage par le groupe IA Responsable.
- Le flou normatif ou son excès compliquent la progression des entreprises sur la voie d'une IA digne de confiance.

CHAPITRE 3

Mettre en œuvre la gouvernance

Pourquoi, quand et comment mettre en place une gouvernance de l'intelligence artificielle dans une organisation ? Sur la base de leurs expériences, les membres du groupe de travail IA Responsable ont croisé leurs regards pour aborder ces questions. Il ressort de leurs réponses un invariable besoin de règles éthiques, des méthodologies de mise en pratique adaptées à chaque entreprise et un objectif majeur : identifier les risques, les prévenir ou en limiter la portée. Autant de constats qui ont permis de lister des recommandations pratiques et des points de vigilance.

Aux origines de la gouvernance

D'où vient le besoin unanimement constaté de s'appuyer sur un corpus de règles éthiques pour encadrer le déploiement de l'IA ? Qu'attendre d'un organe de pilotage et d'un processus de contrôle ? Soumises aux inquiétudes des collaborateurs, aux demandes des clients, au questionnement du monde de la recherche, à la pression des médias, à l'évolution du cadre normatif, voire au respect de leurs engagements RSE (Responsabilité sociétale de l'entreprise), les entreprises n'ont pas le choix. D'autant qu'à ces motifs s'ajoutent les enjeux économiques et concurrentiels du développement de la technologie. Les raisons qui conduisent les organisations sur cette voie sont donc pluridimensionnelles.

En outre, quels que soient leur rôle ou leur fonction au sein de l'entreprise, les acteurs à l'origine des projets de gouvernance se posent eux-mêmes une série de questions de fond. Comment s'assurer que l'utilisation de l'intelligence artificielle ne viendra pas amplifier des risques traditionnels gérés par une entreprise (juridiques, humains, d'image…) ou en générer de nouveaux ? Comment articuler la politique de gouvernance avec la stratégie globale de l'entreprise ? Comment définir précisément le champ d'application de l'intelligence artificielle dans l'organisation ? Comment inclure des dimensions éthiques dans le cycle de vie d'un projet d'IA en complément des critères technologiques ou économiques ? Les règles ou recommandations doivent-elles être globales ou spécifiques selon les usages et les impacts envisagés ? Comment intégrer dans la gouvernance les incertitudes juridiques liées à un champ normatif encore en construction ? Comment impliquer les collaborateurs dans une ambition commune et les rassurer ?

Un cadre a été développé par les coordinateurs du groupe du travail et a été transmis à ses membres afin qu'ils relatent la genèse des approches de la


2P

problématique de gouvernance au sein de leurs structures ; ils ont ensuite confronté leurs expériences desquelles se dégagent quelques points communs.

Historiquement, la communication interne sur le thème de l'intelligence artificielle éthique a débuté il y a moins de cinq ans, souvent à l'initiative des Directions des systèmes d'information (DSI), des services de Recherche et développement (R&D), de l'Innovation ou du Digital. Cette sensibilisation a ciblé prioritairement les membres du Comité exécutif (Comex) et souvent des salariés des branches concernées de l'entreprise. La plupart du temps, ces initiatives isolées ont abouti à de premiers résultats concrets positifs tels que l'identification et la correction de biais ou l'abandon de projets contraires à des principes éthiques. Elles ont également permis de documenter en profondeur la question du moment opportun pour réaliser une évaluation éthique d'une application IA durant son cycle de vie.

Les freins à ces initiatives présentent également quelques similitudes : l'immaturité de la transformation numérique des organisations, que celle-ci soit d'origine technique ou culturelle, la pénurie de collaborateurs spécialisés dans le domaine de l'IA, le caractère secondaire du projet dans la stratégie de l'entreprise[36]. Il faut donc que l'organisation soit suffisamment mûre pour qu'il soit possible de donner un cadre éthique au déploiement de l'intelligence artificielle.

— Retour d'expérience

La direction Digital et Innovation de SNCF Réseau s'est emparée du sujet avec l'intuition que l'IA, dont des solutions commençaient à être testées voire déployées en interne, était de nature à transformer l'activité de l'entreprise, bouleverser l'organisation du travail et questionner la répartition de la valeur.

Quatre ateliers d'une demi-journée ont été organisés rassemblant une vingtaine de personnes issues de différents services du Groupe. Sous des angles différents, chacune de ces réunions débattait d'un projet ou d'une solution d'intelligence artificielle déjà utilisée. Le premier atelier s'est intéressé à Vibrato, une application permettant la maintenance prédictive des voies grâce aux mesures des vibrations, dans sa dimension technique et économique ; le deuxième au projet Mesure sur la charge mentale du point de vue de l'éthique ; le troisième à OpenGov dans une optique RH. Le dernier a synthétisé l'ensemble et bâti des fictions prospectives. Il s'agissait de concevoir un scénario démarrant au moment où survenait un événement important au sein de l'entreprise, à l'horizon d'une dizaine d'années. Grâce à la fiction, les problèmes prenaient corps et devenaient des cas concrets qui soulevaient des questions jusqu'ici ignorées. Des intervenants extérieurs sont venus nourrir les échanges : Giuseppe Longo, mathématicien et épistémologue (ENS, Collège de France)[37], Hubert Etienne, doctorant en éthique de l'intelligence artificielle (ENS, Facebook AI, Oxford)[38] et David Gaborieau, sociologue (CNAM, Université Paris Est)[39]. La synthèse de ces fictions et des questions qu'elles révélaient ont permis de formuler des propositions soumises au Comité exécutif.

Des pratiques diverses

Si le besoin d'une gouvernance est unanimement reconnu par les membres du groupe de travail, ceux-ci font état de pratiques organisationnelles variées. Cette hétérogénéité tient à la diversité des secteurs d'activité, à l'histoire particulière de

36. Deloitte (2020). Thriving in the era of pervasive AI, Deloitte's State of AI in the Enterprise, 3rd Edition.

37. Biographie de Giuseppe Longo

38. Biographie de Hubert Etienne

39. Biographie de David Gaborieau

Si le besoin d'une gouvernance est unanimement reconnu, sa pratique est nécessairement variée.

CETTE HÉTÉROGÉNÉITÉ TIENT À LA DIVERSITÉ DES SECTEURS D'ACTIVITÉ, À L'HISTOIRE PARTICULIÈRE DE CHAQUE ENTREPRISE, À SA TAILLE, À SA CULTURE ET À SES ENJEUX SPÉCIFIQUES.

chaque entreprise, à sa taille, à sa culture, à ses enjeux spécifiques. En la matière, le pragmatisme est donc de rigueur.

Ainsi en va-t-il par exemple de Thales et d'Orange qui l'un comme l'autre ont mis en place une gouvernance mais dont les priorités, champs et modalités d'application s'adaptent aux particularités de leurs enjeux.

— Retour d'expérience

Les solutions proposées par Thales, leader technologique mondial de l'aérospatial, de la défense, de la sécurité et de l'identité numérique, s'appuient de plus en plus sur l'intelligence artificielle. Bien que très nombreux et divers, les domaines concernés présentent une caractéristique identique : l'exigence de fiabilité des systèmes. En effet, l'approximation et les failles sont à l'évidence inacceptables lorsqu'il s'agit de pilotage automatique d'un avion, de sécurisation d'aéroports ou de flux bancaires, de paiement en ligne ou de gestion d'un système d'arme (moyens électroniques permettant aux militaires la réalisation d'une mission et la mise en œuvre autonome d'un armement). Pour ces applications critiques, la notion d'une IA digne de confiance prend, dès lors, tout son sens. Par ailleurs, elles interrogent l'éthique de la technologie et la faisabilité de mise en œuvre de certaines règlementations, le RGPD par exemple. Ainsi, le paiement en ligne ou la reconnaissance faciale à une frontière soulèvent les délicates questions de la protection des données personnelles, de l'infaillibilité des données ou des biais. Ces enjeux inédits ont conduit à la création d'une gouvernance spécifique à l'IA au sein du groupe en 2019. Elle s'appuie sur la « Charte éthique et de transformation numérique » ainsi que sur l'approche Thales TrUE. Celle-ci met en avant une IA transparente, dans laquelle les utilisateurs ont accès aux données exploitées. Elle promeut également une IA intelligible, qui permet d'expliquer et de justifier les résultats, ainsi qu'une IA éthique qui respecte des protocoles standardisés et objectifs, la législation et les droits de l'Homme.

De son côté, Orange[40] entend tenir un rôle clé dans le développement d'une société numérique de confiance. L'opérateur multiservices international voit l'intelligence artificielle comme un tremplin pour accélérer sa propre transformation numérique et pour proposer des services nouveaux à ses clients. C'est pourquoi la question de l'éthique a d'abord été abordée sous l'angle de l'innovation, par la division éponyme, afin d'anticiper le plus en amont possible la gestion des risques, tel que celui des discriminations. La signature d'une charte pour une IA inclusive[41], en avril 2020, a marqué la première concrétisation formelle de l'engagement d'Orange dans une gouvernance en ligne avec ses valeurs et sa promesse sociétale (voir le retour d'expérience p. 60).

PwC[42] a développé pour ses clients une boîte à outils de l'IA responsable qui repose sur cinq piliers, dont la gouvernance. Celle-ci s'entend comme un cadre applicable à tout projet ou utilisation d'IA et pour l'ensemble de l'organisation, de la direction générale aux équipes-projet. Elle doit se concentrer sur les risques inhérents aux projets d'intelligence artificielle et les contrôles à apporter pour les maîtriser tout au long du cycle de vie de la solution d'intelligence artificielle.

Les membres du groupe de travail s'accordent sur la nécessaire création de lieux d'échange, de réflexion et d'arbitrage collectifs entre des experts de domaines divers et des personnes aux fonctions variées au sein de l'entreprise.

4 CONSEILS PRATIQUES POUR AMORCER UNE GOUVERNANCE IA

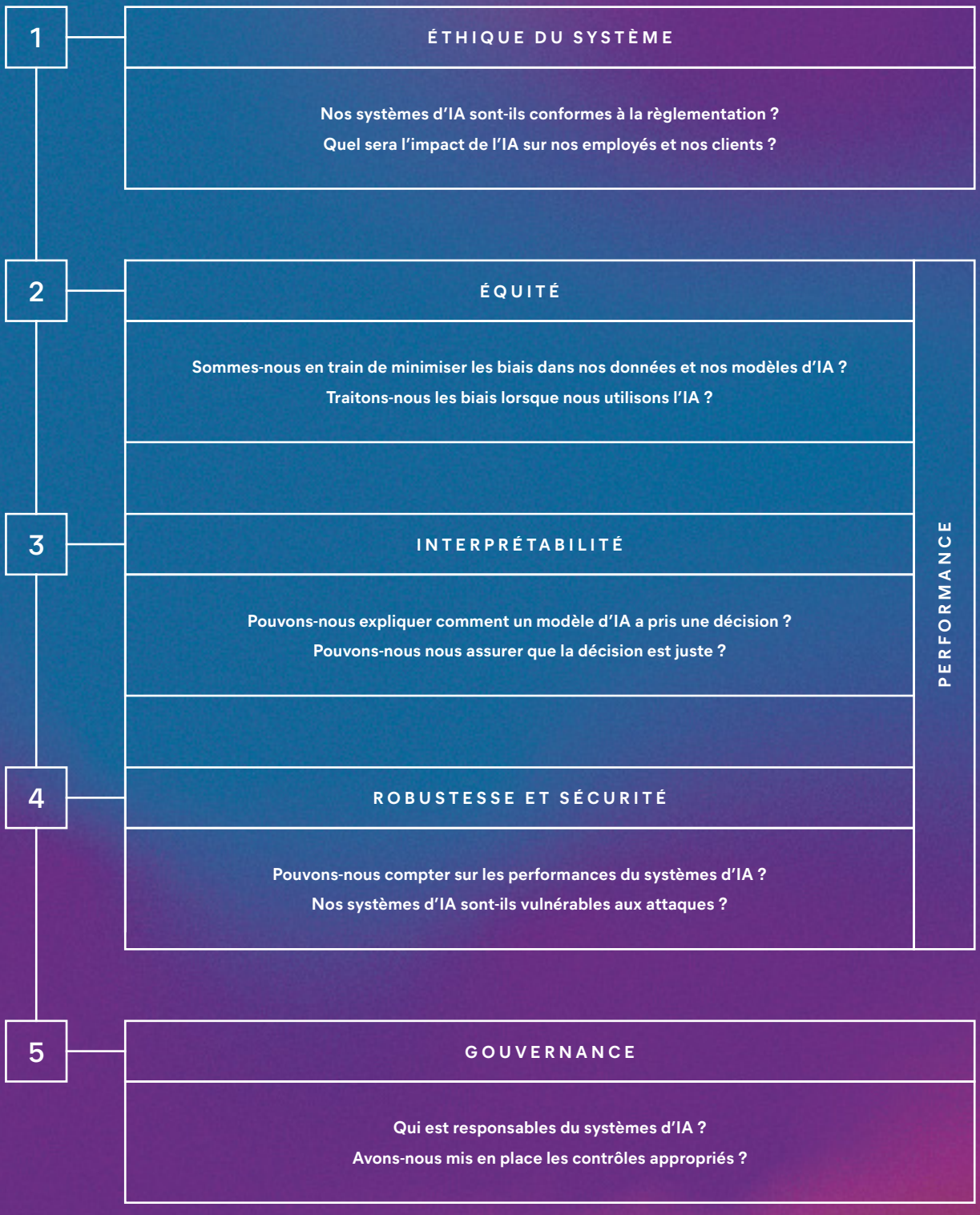
1. Créer un réseau dans les différentes directions de l'entreprises qui utilisent ou sont impactées par des applications d'IA.
2. Nouer des liens avec des chercheurs en intelligence artificielle afin de sensibiliser les collaborateurs avec un discours d'autorité.
3. À travers des ateliers participatifs, créer des scénarii qui soulignent les risques liés à l'IA et leurs enjeux pour l'entreprise.
4. Rédiger une note de diagnostic à destination du Comité exécutif pour informer sur les défis internes et externes de gouvernance de l'intelligence artificielle.

40. Orange, « Orange dévoile sa raison d'être co-construite », 10 décembre 2019

41. Arborus Orange, « Charte internationale pour une I.A. inclusive », avril 2020

42. PwC, étude « 2019 AI Predictions »

<fig.12.> Les cinq piliers de la boîte à outils de PwC pour une IA responsable.



En effet, les solutions à apporter à certaines questions complexes posées par l'utilisation de la technologie, ne peuvent se trouver que dans l'écoute, le débat et la délibération. De plus, cette approche collaborative valorise les différentes parties prenantes qui envisagent ainsi d'un œil plus confiant les usages de l'IA. Toutefois, là encore, l'articulation des organes qui poursuivent cet objectif diffèrent d'une organisation à l'autre (voir p. 74).

Identifier les enjeux

Avant de circonscrire le périmètre de la gouvernance de l'IA en entreprise ainsi que ses outils et protocoles, le groupe de travail a défini précisément ses enjeux et ses objectifs.

Chaque structure doit se focaliser en priorité sur les risques qu'elle distingue clairement et désigne comme des défis majeurs. Sa gouvernance vise donc à concevoir et mettre en œuvre des garde-fous adaptés. Les dangers liés à l'intelligence artificielle ne lui sont que partiellement imputables : biais des données et des modèles, opacité dans la prise de décision, critères d'apprentissage et de calcul antagonistes.

Mais la particularité de cette technologie tient à ce qu'elle est aussi susceptible d'amplifier les risques qu'affronte et gère ordinairement toute entreprise — stratégiques, organisationnels, humains, juridiques… De nombreux membres du groupe de travail ont témoigné qu'ils prenaient appui sur le dispositif de gouvernance de l'IA pour corriger ou atténuer ce type de problèmes grâce à des outils, notamment techniques, et des protocoles.

Il apparaît ainsi que la gouvernance de l'intelligence artificielle consiste prioritairement à identifier les dommages potentiels liés à la conception et à l'utilisation de cette technologie afin de les prévenir, les amender ou en limiter la portée.

Avant de se lancer, quelques lignes directrices

Il s'agit d'abord de définir et formaliser des principes et valeurs structurants qui détermineront la stratégie, organiseront la démarche et serviront de base à une déclinaison opérationnelle, notamment en hiérarchisant les questions à traiter par l'entreprise.

Après avoir recensé les éventuelles instances de gouvernance et réflexions préexistantes et y avoir intégré le pilotage de la politique de l'éthique de l'IA, il devient possible de mettre sur pied une méthodologie de management des risques éthiques, légaux et de robustesse dans les projets ou processus ; celle-ci peut, selon les cas, aller jusqu'aux procédures d'arbitrage sur des questions de fond, que ce soit par le biais d'un comité décisionnaire ou en s'appuyant sur une certification, entre autres (voir p. 74).

Dans la mise en œuvre, les responsables du pilotage de l'éthique dans l'intelligence artificielle doivent veiller à adopter une posture de facilitateur. L'IA offre en effet des opportunités techniques, sociales, sociétales et économiques formidables, elle ouvre des possibilités inaccessibles jusqu'ici. Sa gouvernance a vocation à stimuler son développement en intégrant les autres enjeux de l'entreprise, qu'ils soient d'ordre réglementaire, éthique, juridique, etc. Ce comportement s'avère d'autant plus cardinal que toute attitude de censeur provoquerait un rejet de la part des collaborateurs alors que leur adhésion constitue l'une des clés de la réussite du projet. Ainsi, organiser l'acculturation et la formation à l'éthique — voire les recrute-

ments nécessaires — en particulier dans les équipes chargées de concevoir et mettre en œuvre des outils utilisant l'intelligence artificielle représente un enjeu primordial pour le succès de la gouvernance. Au-delà, cela doit s'accompagner d'une diffusion dans toute l'organisation d'une véritable « culture de l'éthique » (voir p. 80).

À cet égard, documenter et communiquer les décisions concernant les choix de gouvernance ressort comme un point important. C'est en effet un moyen efficace d'améliorer la compréhension tant des questions qui se posent que des réponses possibles, de prouver la volonté de transparence, de rassurer les parties prenantes et de faciliter les audits à venir. Ainsi, on pourra par exemple expliquer les raisons et les méthodes qui président à la mise en place d'un comité d'éthique.

Enfin, une veille juridique, industrielle et académique constante s'avère fructueuse car les nombreux acteurs travaillant sur la gouvernance proposent régulièrement de nouveaux outils et méthodes, potentiellement utiles à la mise en place et au pilotage d'une intelligence artificielle digne de confiance.

Un sponsoring puissant et actif

L'engagement fort et sans ambiguïté des plus hautes autorités de l'entreprise conditionne la réussite de la politique de gouvernance. Les membres du groupe de travail ont par exemple constaté dans leurs expériences que son absence constituait un frein aux premières initiatives, au moment où le besoin émergeait.

La participation d'un ou plusieurs membres du Comité exécutif (Comex) aux organes nouvellement créés représente un signal fort qui confère une légitimité à la mise en place d'une gouvernance de l'IA au sein de l'organisation.

En effet, cette implication active matérialise la vision que porte la direction générale sur les enjeux liés à la technologie (risques et opportunités) et symbolise l'importance accordée aux règles d'éthique ainsi qu'à leur respect dans l'entreprise. L'intégration du développement de l'intelligence artificielle au plan stratégique témoigne également tout à la fois de la portée du sujet et de son caractère transversal : l'intelligence artificielle n'est pas l'affaire de tel ou tel département ou direction de l'entreprise mais elle requiert la mobilisation de tous les collaborateurs. Le financement et la validation d'une charte éthique peuvent traduire cet engagement.

3 CONSEILS PRATIQUES POUR FACILITER LE SPONSORING DU COMEX

1. Sensibiliser le Comex par l'intervention d'un ou plusieurs experts qui témoignent des enjeux business de la gouvernance de l'IA.
2. Anticiper les inquiétudes du Comex en préparant des réponses à ses questions éventuelles.
3. Présenter au Comex une feuille de route, complémentaire du diagnostic, clarifiant les jalons-clés de la gouvernance IA : intégration au plan stratégique, validation d'une charte, création d'un comité, etc.

— Retour d'expérience

À la MAIF, deux directeurs généraux adjoints, dans leurs domaines respectifs, sont devenus sponsors des projets internes portant sur les données et l'IA : la direction de la Sécurité des Systèmes d'Information et la direction de l'Assurance de la personne, de la data & de l'actuariat produit. Ces questions sont partie intégrante de leurs responsabilités et ils sont porteurs de ces sujets au sein du Comité de direction générale. Les engagements de la MAIF sont formalisés dans la « Charte pour un monde numérique résolument humain et éthique », publiée en 2017.

Des organes aux acteurs : la gouvernance au quotidien

Globalement, le dispositif de gouvernance doit définir les objectifs de la politique pour une IA digne de confiance et les plans d'action pour les atteindre. Il détermine en conséquence des normes internes et des processus de contrôle de conformité

L'intelligence artificielle n'est pas l'affaire de tel ou tel département

OU DIRECTION DE L'ENTREPRISE MAIS ELLE REQUIERT LA MOBILISATION DE TOUS LES COLLABORATEURS.

adaptés à l'organisation. Il est ainsi important de le construire en cohérence avec les fonctionnements et les pratiques managériales habituels de l'entreprise. L'expérience des membres montre en effet que, si le but reste identique, les méthodes de mise en œuvre diffèrent.

Les commissions ou comités déjà chargés de la gestion des risques ou au sein d'une organisation se saisissent parfois de ce nouveau thème sous son angle éthique et l'intègrent dans leurs procédures. Il arrive également que des instances intéressées par la seule problématique des données ou de l'innovation élargissent naturellement leur périmètre à l'intelligence artificielle. Dans d'autres cas, certaines entreprises déjà dotées d'une direction de l'éthique et de la déontologie dont le champ de responsabilité s'étend bien au-delà de l'IA (par exemple aux problèmes rencontrés individuellement par les collaborateurs dans le cadre de leur travail) peuvent aussi inclure l'intelligence artificielle dans leur mission. Certaines études pointent vers le département des Ressources humaines comme initiateur de la démarche. Enfin, des structures *ad hoc* sont quelquefois créées de toutes pièces. L'ambition reste néanmoins la même dans tous les cas de figure : donner un cadre d'action afin de définir des principes communs applicables dans toute l'organisation.

Les périmètres de responsabilité des organes de gouvernance ainsi que leur composition varient également selon les spécificités organisationnelles des entreprises. Ainsi, tel groupe décidera que les sujets relevant de l'éthique de l'IA seront traités de façon centralisée tandis que tel autre délèguera la définition des standards d'évaluation à ses départements ou filiales concernés. De même, les organisations sollicitent parfois des personnalités extérieures venues, selon les cas, de l'industrie technologique, des sciences humaines et sociales, des ressources humaines ou du droit pour intégrer leur comité de gouvernance. *A contrario,* d'autres décident de se reposer sur des expertises internes.

Mais quelles que soient leur forme, invariablement ces organes articulent les dimensions stratégique et opérationnelle.

Les membres du groupe de travail IA Responsable ont listé les prérogatives de chacune de ces instances potentielles, puis indiqué les personnages-clés qui y participent ainsi que leur rôle.

— S'APPUYER SUR UN COMITÉ ÉTHIQUE OU DE GESTION DES RISQUES EXISTANTS :

Cet organe est capable de garder une hauteur de vue sur les grandes orientations stratégiques de l'entreprise et dans ce cas, il est pertinent d'ajouter l'intelligence artificielle à l'ensemble de son périmètre de responsabilités. À travers ses réflexions et son action, il porte un regard synoptique sur les enjeux de l'éthique appliqués à l'intelligence artificielle au sein de l'organisation. Il s'intéresse à tous les départements et activités de l'entreprise et prend en compte les conséquences du développement de la technologie dans ses aspects économiques, technologiques, sociaux, organisationnels et d'image.

Dans ce cadre, le comité sera par exemple garant de la « Charte éthique de l'IA », validée par la direction générale, incluant un code déontologique qui encadre la mise en œuvre de solutions d'intelligence artificielle dans les différents départements de l'entreprise au regard de ce que la Charte précise (voir p. 74). Il s'informera aussi des besoins qui apparaissent au fur et à mesure du temps, capitalisera les retours d'expérience sur les usages, publiera des avis, élaborera des préconisations générales et proposera des modalités d'accompagnement aux projets et départements de l'entreprise.

À la tête de ce Comité, un responsable anime les débats de représentants des fonctions principales de l'entreprise : RH, SI, R&D, Distribution, Gestion / Production, RSE, voire expert externe.

Donner un cadre d'action pour définir des principes applicables dans toute l'organisation.

TELLE EST L'AMBITION

DANS TOUS LES CAS DE FIGURE.

Cette approche basée sur un comité éthique ou de gestion des risques n'est pas la seule possible ; dans les entreprises ne possédant pas cet organe de gouvernance, le Comité de direction générale peut aussi assurer les fonctions d'arbitrage sur des sujets éthiques, en gardant la hauteur de vue nécessaire à la prise décision.

— METTRE EN PLACE UN ORGANE "ÉTHIQUE & IA" DÉDIÉ

Une approche plus agile et au cœur des projets peut s'envisager. On peut proposer par exemple la mise en place de la gouvernance éthique de l'IA autour d'une fonction-clé (qu'on pourrait nommer Chief Ethics Officer). Celui-ci serait accompagné d'un collectif d'experts des données et de l'intelligence artificielle, capable d'une part d'évaluer les enjeux de l'IA au regard des valeurs et des objectifs de l'entreprise.

Les missions de ce collectif d'experts ont une vocation opérationnelle :

Auprès de l'organe de gouvernance décisionnaire des questions liées à l'éthique de l'IA (Comité éthique élargi à l'IA, Comité de direction générale…)
– Rédiger une « Charte éthique de l'IA » incluant un code déontologique qui règlemente la mise en œuvre de solutions d'intelligence artificielle dans les différents départements de l'entreprise, conforme aux objectifs éthiques poursuivis ;
– Déployer un radar des projets, produits ou utilisations basés sur l'IA au sein de l'entreprise et être garant du suivi de leur évaluation, au regard des engagements pris dans la Charte ;
– Alerter et mettre à contribution l'organe sur des sujets porteurs de nouvelles questions éthiques. Informer sur les enjeux et impacts afin de faciliter les arbitrages.

Au niveau des projets et auprès des départements de l'entreprise
– Concevoir et réaliser des protocoles et outils qui permettent d'appliquer sur le terrain les règles éthiques de la Charte et ce quel que soit le département concerné dans l'entreprise : liste des vérifications à faire (check-lists) à tous les stades du développement d'une solution d'IA, socles technologiques transverses, référentiel de compétences…
– Acculturer et former les métiers aux enjeux de cette technologie (opportunités et impacts)
– Recenser les besoins et modalités d'accompagnement dans les différents départements de l'entreprise : nouveaux métiers ou évolution de métiers, formations, nomination de « Champions IA »…

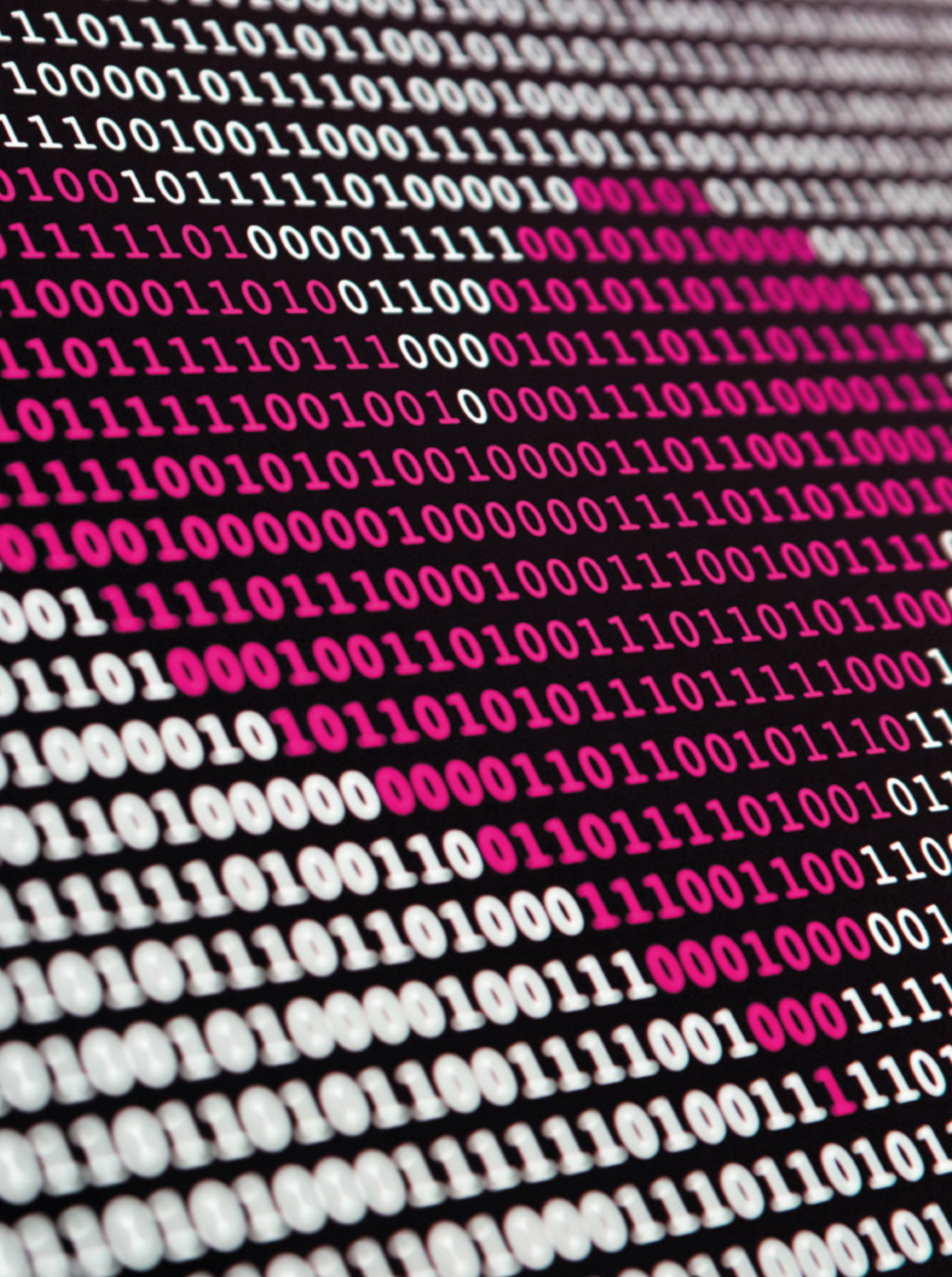
Sur l'ensemble de ces missions, le collectif s'accompagne d'un réseau de contributeurs internes (RH, Marketing, Gouvernance projets…) et externes afin de construire ensemble le référentiel de valeurs, éclairer les décisions et inscrire concrètement les projets dans cette démarche d'éthique *by design.*

— Retour d'expérience

Afin de mettre en place sa « Charte éthique et transformation numérique », Thales a constitué un groupe de travail où sont représentées les principales fonctions de l'entreprise : Recherche & Technologie, Opérations, Ressources humaines, Communication, Éthique et Responsabilité d'entreprise, Juridique et contrat, Commercial… Des collaborateurs venus de l'international ont également intégré cette équipe, un apport crucial qui permet de prendre en compte des différences culturelles régionales. Sur des questions de société aussi fondamentales

7 CONSEILS PRATIQUES POUR UN FONCTIONNEMENT OPTIMAL DES ORGANES DE GOUVERNANCE

1. Recenser les comités préexistants au sein de l'organisation et éviter les doublons.
2. Pour débuter, créer une structure légère afin d'informer, partager et sensibiliser les collaborateurs sur les risques liés à l'IA. À ce stade, la mise en place de comités décisionnels plus importants, pourrait être perçue comme une lourdeur administrative par les collaborateurs.
3. Choisir des membres suffisamment au fait de la technologie pour siéger dans les comités, afin qu'ils nourrissent une réflexion pertinente, émettent des avis circonstanciés et décident de façon éclairée.
4. Être attentif à l'homogénéité des niveaux hiérarchiques représentés dans les comités (ou de reconnaissance au sein des organisations « à échelle double ») pour que toute prise de parole apparaisse légitime et soit écoutée.
5. Veiller à l'assiduité aux réunions dont la richesse tient à l'échange et à la diversité des points de vue.
6. Installer et partager des valeurs communes d'écoute, de droit d'alerte, de liberté d'expression ; prévoir aussi des temps de débats contradictoires.
7. Maintenir la motivation et la cohésion des membres en créant un esprit de cohésion : animation de la communauté par des échanges informels réguliers, organisation des réunions à un rythme cadencé.



que la diversité, les libertés individuelles ou la liberté d'expression, la disparité des perceptions est très forte, révélant parfois des antagonismes profonds. Ce groupe de travail exclusivement composé de collaborateurs, fait néanmoins appel, chaque fois qu'il le juge nécessaire, à des experts extérieurs qui viennent partager leurs réflexions.

Enfin, sur la base des retours d'expérience des équipes de Thales, ce groupe de travail a également vocation à réviser régulièrement les principes de la « Charte éthique et transformation numérique » afin de maintenir leur pertinence et leur adéquation aux enjeux éthiques associés à la transformation numérique.

Embarquer les collaborateurs dans le projet

La sensibilisation en interne, à tous les niveaux hiérarchiques, représente un point décisif. Au-delà de la direction générale, les managers opérationnels, les équipes de conception et de développement et les collaborateurs-utilisateurs doivent percevoir les opportunités de cette technologie sans en ignorer les risques. Ne serait-ce que parce que l'intelligence artificielle transforme en profondeur l'entreprise et ses métiers, les salariés sont les premiers concernés et ont besoin de repères clairs[42]. Démystifier l'intelligence artificielle, tant ses aspects positifs que négatifs, la rend accessible et facilite son appropriation.

Ainsi, associer à la réflexion sur les premiers jalons de la gouvernance des représentants de toutes les entités de l'entreprise pour recueillir leurs suggestions ou remarques déclenche bien souvent le processus d'acculturation : il n'est de meilleur ambassadeur que celui dont émane l'idée.

Cette appropriation, progressive, relève d'un travail conjoint des Ressources humaines et de la Communication interne.

La promotion du modèle de gouvernance ciblera en particulier les managers opérationnels, clefs de voûte de la mise en œuvre de la stratégie. Selon les pratiques de communication usuelles de l'entreprise, cela pourra prendre la forme de séminaires, de formations spécifiques ou bien encore s'intégrer dans des programmes plus généraux sur l'éthique et la politique de RSE de l'entreprise. Le rôle des Comités et leurs procédures sera expliqué et valorisé auprès des directeurs, managers de terrain et chefs de projet.

Au niveau global, un dispositif de formation « maison » destiné à tous les salariés et aux nouveaux entrants donnera une vision synoptique des enjeux de l'IA éthique en la resituant dans le contexte spécifique de l'organisation. Les modules devront expliquer les principes généraux ainsi que les organes et outils de gouvernance et compléter cet exposé par une application à des cas concrets. Il est également judicieux de valoriser des expériences de reconversions professionnelles grâce à l'intelligence artificielle. À l'échelle des équipes de conception, les procédures d'accueil des nouveaux collaborateurs peuvent inclure une « Charte du Data scientist ».

La Communication interne pour sa part poursuivra des objectifs corrélés : informer et expliquer, susciter l'adhésion en donnant du sens au projet, valoriser les expériences positives et leurs acteurs vertueux. Il est primordial de communiquer sur la gouvernance en la replaçant dans le tableau général de la stratégie de déploiement de l'IA dans l'entreprise. Le but est en premier lieu de relayer les convictions et valeurs de l'organisation sur l'éthique de l'intelligence artificielle et d'informer les collaboratrices et collaborateurs des instances et procédures mises en place.

Cette pédagogie doit aussi les alerter sur les opportunités offertes par la technologie et les risques inhérents à son utilisation, en toute transparence, afin de donner du sens à la stratégie et de créer les conditions de la confiance. En

5 CONSEILS PRATIQUES POUR FAVORISER L'ACCULTURATION EN INTERNE

1. Étape initiale vers l'appropriation, associer aux premières réflexions sur la gouvernance des représentants de tous les départements de l'entreprise concernés par l'IA pour recueillir leurs suggestions.
2. Promouvoir particulièrement la démarche de gouvernance auprès du public des managers opérationnels, interlocuteurs naturels des collaboratrices et collaborateurs sur le terrain et garants de la mise en œuvre de la stratégie.
3. Créer des dispositifs de formation pour donner une vision générale des enjeux éthiques de l'IA et des réponses apportées par l'entreprise.
4. Communiquer et sensibiliser en interne sur les valeurs de l'éthique de l'intelligence artificielle vues par l'entreprise, sur les opportunités et les risques, pour donner du sens et susciter la confiance et l'adhésion.
5. Informer des raisons de la création d'une gouvernance de l'IA, de ses objectifs, de ses modes opératoires et des gains attendus afin de donner les moyens de passer à l'action sur le terrain.

42. Deloitte (2020). Talent and workforce effects in the age of AI, Insights from Deloitte's State of AI in the Enterprise, 2nd Edition.

La sensibilisation en interne, à tous les niveaux hiérarchiques, représente un point décisif.

AU-DELÀ DE LA DIRECTION GÉNÉRALE, LES MANAGERS OPÉRATIONNELS, LES ÉQUIPES DE CONCEPTION ET DE DÉVELOPPEMENT ET LES COLLABORATEURS-UTILISATEURS DOIVENT PERCEVOIR LES OPPORTUNITÉS DE CETTE TECHNOLOGIE SANS EN IGNORER LES RISQUES.

réitérant le propos de la « formation maison », la communication interne s'assure que le message est diffusé jusqu'au plus près du terrain. Cette redondance a également pour fonction de marquer les esprits, de faciliter la mémorisation et d'afficher clairement l'importance accordée au sujet par la direction générale.

Pour remplir cette mission, tous les supports classiques de communication interne peuvent être mobilisés selon les ressources et habitudes de l'organisation, depuis les discours de membres du Comex jusqu'à l'intranet ou aux groupes collaboratifs sur un réseau social d'entreprise.

Enfin, les actions de communication valoriseront les choix vertueux afin de les transformer en avantages compétitifs et de rendre compte des bénéfices de la politique de gouvernance. La mise en avant de projets et produits à succès mais aussi les parcours de collaborateurs enrichis par l'IA et sa dimension éthique viseront à susciter l'engagement.

Des outils pour agir

Au-delà de la définition d'une vision stratégique, la gouvernance se doit de concevoir, mettre au point et diffuser des ressources, c'est-à-dire des outils qui architectureront et guideront le travail quotidien des équipes opérationnelles.

Première étape, la rédaction d'une « Charte d'éthique de l'IA ». C'est elle qui, en mettant en lumière les valeurs de l'organisation et ses principes intangibles, les défis en jeu et les objectifs poursuivis, donne du sens à la gouvernance.

Elle constitue un cadre fort parce qu'elle donne des lignes de conduite, mais suffisamment souple pour permettre d'amorcer l'intégration de l'éthique dans les processus métiers. Se voulant concrète, elle inclut un code déontologique qui règlemente la mise en œuvre de solutions d'intelligence artificielle dans les différentes structures de l'entreprise et s'enrichit au fil de l'eau par les Retour d'expérience ou les nouveaux cas d'usages.

Pour ce qui est des instruments opérationnels, les check-lists de vérification du respect des règles éthiques à tous les stades de la vie d'un modèle d'intelligence artificielle semblent avoir apporté la preuve de leur efficacité en dépit des difficultés rencontrées (voir encadré ci-dessous). Elles présentent en effet le grand avantage d'être simples à utiliser. Néanmoins, il sera instructif de vérifier cette efficience dans le temps en mesurant leurs impacts dans les années qui viennent.

— Retour d'expérience

Orange a déjà de nombreux processus mis en place pour la gestion et l'anticipation des risques (pour la sécurité, le RGPD par exemple). Il s'agit donc d'ajouter un volet supplémentaire, spécifique aux enjeux éthiques de l'IA, cohérent et compatible avec l'existant. Par ailleurs, il convient de trouver un équilibre entre l'exhaustivité des questions dans la check-list et l'efficacité opérationnelle. De la même façon, les vérifications ne peuvent être monolithiques : elles doivent être adaptées au niveau de maturité des projets, entre recherche, anticipation et mise en service par exemple.

Par ailleurs, le partage des responsabilités doit être clair : si les personnes en charge des questions éthiques accompagnent les projets à leurs débuts, il revient à chaque équipe de développement d'un système d'IA de trouver les solutions adaptées à son contexte.

Reste également la question de l'application de la check-list à tous les projets ou services déjà opérationnels en tenant compte de leur diversité : par exemple du chatbot à l'optimisation réseau.

Du côté du Cercle InterElles, les check-lists ont pour fonction d'évaluer le niveau de maturité d'un projet d'IA ou d'une entreprise sur les discriminations de genre afin d'identifier des priorités et de mettre au point un plan d'action. Avant la généralisation de la check-list à l'ensemble d'une organisation, il est judicieux de tester le questionnaire auprès d'un panel limité de personnes concernées pour vérifier leur compréhension de la formulation des questions. Dans le même ordre d'idée, il est nécessaire, à l'instar des pratiques des sondeurs, de mesurer le temps qu'y consacrera le répondant afin de ne pas le rebuter par une analyse trop longue de ses pratiques. Le choix du ou des répondant(s) est également important afin d'avoir le bon niveau de compréhension des projets et des enjeux liés à l'intelligence artificielle. À ce titre, privilégier un mode collectif (équipe-projet…) pour utiliser une check-list permet d'éviter les automatismes voire… l'autosatisfaction dans les réponses. D'autant que se poser collectivement la question de sa maturité en termes d'IA aide à l'acculturation des collaborateurs sur ce sujet.

L'intégration de l'éthique dans les critères de pilotage des risques constitue un point de référence pour les opérationnels. C'est la raison pour laquelle les équipes de concepteurs ou d'utilisateurs de solutions basées sur l'IA doivent souvent adapter leurs méthodes de conduite de projet pour y introduire la dimension éthique. L'objectif consiste à intégrer aux process des phases de questionnement sur la conformité aux règles internes de gouvernance, par exemple pour chercher des biais éventuels dans les données au cours de l'apprentissage des algorithmes de machine learning. L'intégration des processus d'évaluation de l'IA dans le corpus des procédures « métier » existantes facilitera leur mise en place et leur application tout en limitant leur complexité de mise en œuvre.

À l'expérience, la réalisation d'un glossaire spécifique à chaque projet peut s'avérer efficace en phase de conception afin d'éviter toute ambiguïté et que chacun parle le même langage. Dans le cadre d'une démarche de partage des bonnes pratiques, il est intéressant de valoriser dans toute l'entreprise ce qui a fait la réussite d'un projet. Par exemple, montrer qu'une check-list « type » de vérification des points-clés a évolué au fil des retours d'expérience pour mieux traiter les problématiques telles que les biais, les impacts humains et organisationnels d'une solution, le respect du RGPD… À cette panoplie de procédures peuvent s'ajouter d'autres protocoles plus techniques qui éclairent la prévisibilité des algorithmes en termes de biais discriminatoires notamment (voir p. 60).

De même qu'ils ont permis la création de filières dans les entreprises quand ont émergé de nouveaux thèmes stratégiques tels que la construction durable ou l'alimentation responsable, les référentiels conçus par les organismes extérieurs jouent un rôle majeur pour asseoir la politique de gouvernance de l'IA responsable. Ce sont en effet des instruments solides qui structurent la démarche, l'objectivent et la rendent opposable. Trois possibilités, éventuellement combinées, s'offrent aux entreprises dans ce domaine : les labels volontaires, les certificats et les normes.

Dans la première configuration, un acteur propose un référentiel de pratiques responsables aux entreprises qui choisissent de le suivre. Citons pour exemple la charte Arborus, choisie par Orange (voir p. 60) ou Substra Foundation, organisation non lucrative qui travaille au développement d'écosystèmes de data de confiance, qui soutient les entreprises qui souhaitent autoévaluer leur démarche (voir encadré).

— Retour d'expérience

Depuis mi-2019, Substra Foundation anime une démarche participative dans le but de concevoir un référentiel « data responsable et de confiance » dont les organisations se saisissent pour apprécier la

L'intégration de l'éthique constitue un point de référence

DANS LES CRITÈRES DE PILOTAGE

DES RISQUES POUR LES OPÉRATIONNELS.

maturité de leur politique d'éthique de l'IA et pour progresser en puisant dans des ressources techniques adaptées à chaque item d'évaluation. Dreamquark, qui propose des solutions innovantes d'analyse de données aux secteurs de la banque et de l'assurance, s'est appuyé sur ce système pour autoévaluer ses modèles d'intelligence artificielle à toutes les étapes de leur cycle de vie. La startup a ainsi repéré les domaines dans lesquels elle était bien avancée et identifié les actions à mettre en œuvre pour s'améliorer dans d'autres.

L'entreprise peut aussi faire le choix de faire évaluer ses actions en faveur de l'IA digne de confiance à l'aune du référentiel d'un certificateur extérieur accrédité par le Cofrac (Comité français de certification), qui, sur la base d'un audit, délivre des certificats officiels de conformité. Le Laboratoire National de Métrologie et d'Essai (LNE) vient à ce titre de lancer un groupe de travail avec une quinzaine d'entreprises.

Dernier cas de figure, l'organisation peut décider d'observer des règles définies par un comité de normalisation national, européen ou international. Des projets sont notamment en cours au CEN-CENELEC (groupe de travail dans lequel la France est représentée par l'Afnor (Association française de normalisation), l'IEEE, l'IEC ou encore l'ISO.

2 CONSEILS PRATIQUES POUR RENFORCER L'EFFICACITÉ DES OUTILS

1. Tester les check-lists afin d'en évaluer la pertinence et les adapter si nécessaire avant de les intégrer au processus d'évaluation finale.
2. Rassembler dans un même document les situations à risque identifiées, les comportements à adopter en cas de non-conformité, les outils mis à disposition. Cette synthèse évoluera chaque fois que se présenteront de nouvelles problématiques et les solutions mises en œuvre.

Évaluer la gouvernance pour rester efficace dans la durée

La crédibilité et l'utilité de la gouvernance IA, tant en interne qu'en externe, dépendent de l'évaluation de son efficacité dans le temps. Cela permet d'ajuster son fonctionnement le cas échéant. Quels moyens pour cela ? À ce stade, trois pistes se dégagent au sein des organisations. D'abord, le Comité IA doit définir et mesurer les performances des recommandations d'améliorations éthiques des projets d'intelligence artificielle. Ensuite, une enquête auprès des managers et des collaborateurs évaluera les impacts, positifs ou négatifs, de cette gouvernance sur leurs activités. Selon ses conclusions, il conviendra de modifier ou non les protocoles de *design,* développement, test et déploiement des applications d'intelligence artificielle. Pour finir, des audits réguliers jaugeront son degré d'appropriation en interne.

3 CONSEILS PRATIQUES POUR UN DISPOSITIF D'ÉVALUATION PERFORMANT

1. Faire en sorte que le responsable du Comité IA puisse faire remonter les points d'attention et les décisions prises directement au Comex, en priorisant les sujets.
2. Le cas échéant, utiliser le dispositif d'écoute interne existant pour recueillir l'opinion des managers et des collaborateurs sur les impacts de la gouvernance IA.
3. Convenir avec l'équipe d'audit interne d'un cadre de vérification facilement utilisable et planifier les audits à venir.

Faire rayonner sa gouvernance à l'externe

Outre la partie ciblant l'interne qui vise à favoriser l'acculturation des collaborateurs, la communication comprendra un volet externe, afin de mettre en lumière les actions de l'organisation en faveur de l'IA responsable et d'informer les différents acteurs de son écosystème de l'avancement de cette politique. Elle viendra en somme enrichir l'image corporate de l'entreprise en s'appuyant sur différents supports et initiatives. D'abord une rubrique du rapport d'activité sera consacrée à l'action des comités et éventuellement aux KPIs *(Key Performance Indicators)*, rubrique qui rendra compte des évolutions sur le sujet d'année en année. Par ailleurs, des interventions de responsables de la politique de gouvernance lors de manifestations d'organisations professionnelles (Syntec, Agora, Medef…) ou au cours de salons professionnels du type VivaTech ou encore Big Data Paris serviront de levier pour faire rayonner les initiatives de l'organisation. Ceci constituera en outre un bon moyen d'échange avec d'autres entreprises confrontées à des questionnements similaires et permettra de comparer les pratiques.

À RETENIR

- Le besoin de gouvernance de l'IA est universel mais il n'existe pas de modèle unique de mise en œuvre. Celle-ci prend des formes diverses, adaptées aux enjeux singuliers de chaque entreprise, à son secteur d'activité et à sa taille, ainsi qu'à son histoire et sa culture.
- La gouvernance consiste prioritairement à identifier, prévenir et corriger les potentiels effets délétères de l'utilisation de l'intelligence artificielle ou à contenir l'intensification des risques traditionnels de l'entreprise qu'elle génère.
- Aucune gouvernance n'est possible sans un soutien explicite des plus hautes instances de l'organisation.
- L'intégration précoce du projet de gouvernance de l'IA aux structures comparables déjà en place dans l'organisation — Comité éthique, par exemple — constitue un facteur de succès.
- L'acculturation des collaborateurs est une condition sine qua non de la réussite. On s'appuiera sur l'action conjointe des Ressources humaines et de la Communication interne pour créer un climat de confiance.
- Les principaux instruments opérationnels de gouvernance sont la Charte éthique de l'IA, les check-lists de vérification et de conformité ainsi que les référentiels conçus par des organismes externes à l'entreprise.
- La stratégie de la gouvernance est édictée par un Comité éthique étendu à l'IA ; un ou plusieurs autres organes se chargent de sa mise en œuvre opérationnelle.
- Il est essentiel de mesurer les progrès et les points à améliorer grâce à des audits réguliers.
- La communication externe permet de mettre en valeur les initiatives de l'entreprise et d'échanger avec des pairs.

4 CHAPITRE

Autoévaluer sa gouvernance

S'il est trop tôt pour établir un bilan fiable de la gouvernance de l'IA en France étant donné le caractère encore évolutif du sujet, ce guide se propose néanmoins de fournir un instrument d'appréciation de sa progression.

LES TRAVAUX DU GROUPE IA RESPONSABLE ont en effet permis d'identifier des pratiques puis de mesurer l'avancement de leur mise en œuvre dans les organisations. Ce canvas associe les démarches à adopter pour une bonne gouvernance et cinq niveaux d'implémentation. Le premier correspond à une première approche de la problématique de l'IA digne de confiance : compréhension des concepts, du contexte règlementaire, volonté d'implémenter des solutions d'intelligence artificielle ; le cinquième représente la forme la plus aboutie de la gouvernance. Pour chaque niveau de maturité, il est utile de mener des évaluations (ou audits) sur le caractère responsable des solutions d'intelligence artificielle de l'organisation ainsi que des procédés et instances de gouvernance de l'IA mis en place. Pourvu qu'elle s'en donne les moyens financiers, organisationnels et humains, toute entreprise est en mesure de déployer ses systèmes, produits et services d'intelligence artificielle dans le respect de principes d'éthique.

Comme toutes les initiatives d'Impact AI, cette méthode de mesure du niveau de maturité se veut concrète et utile. Elle est cependant dépendante du contexte général (réglementaire, commercial, politique…) de l'organisation qui l'emploie et doit être modifiée en fonction de l'évolution de celui-ci. Il est ainsi important d'adapter l'échelle d'appréciation.

Cette échelle est ici présentée afin que les organisations qui le désirent s'en saisissent et évaluent elles-mêmes ainsi leur propre dispositif de gouvernance. À titre indicatif, en moyenne, les membres du collectif Impact AI se situent au niveau 3.

Les prises de position des autorités françaises et européennes favoriseront l'approfondissement de la réflexion et modifieront peut-être les actions déjà en place dans les organisations. Les catégories d'évaluation ici présentées sont amenées à évoluer avec le développement de l'intelligence artificielle au sein des entreprises en France. Il semble ainsi essentiel de répéter l'exercice mené dans ce guide dans les années à venir afin de caractériser en détail l'avancement de la gouvernance de l'IA en France.

Nous invitons nos lecteurs à partager avec Impact AI (contact@impact-ai.fr) leurs expériences sur le sujet, à l'aide des critères identifiés ci-dessus, afin d'enrichir le guide et spécialement l'échelle de maturité.

Agnes Van de Walle
— PRÉSIDENTE D'IMPACT AI
DIRECTEUR DIVISION
PARTENAIRES ET STARTUPS
MICROSOFT FRANCE

ÉCHELLE DE MATURITÉ DE LA GOUVERNANCE IA SELON IMPACT AI

THÈME	NIVEAU 1	NIVEAU 2	NIVEAU 3	NIVEAU 4	NIVEAU 5
PRINCIPES ÉTHIQUES	Recensement et comparaison des principes éthiques des autorités (UE, OCDE...)	Positionnement des principes éthiques internes de l'organisation par rapport à ceux définis par les autorités.	Publication en interne des principes éthiques de l'IA (charte...).	Publication à l'externe des principes éthiques de l'IA.	Audit de l'application des principes éthiques listés dans la charte.
SPONSORING	Aucun responsable n'est identifié. Les collaborateurs s'organisent spontanément pour explorer le sujet et faire remonter le besoin au Comex.	Le Comex confirme son sponsoring et nomme une collaboratrice ou un collaborateur responsable du développement d'une gouvernance de l'IA.	Une personne responsable de la gouvernance de l'IA a été nommée et agit sans disposer toutefois de toute la légitimité ou des ressources nécessaires pour mobiliser l'ensemble des parties prenantes pertinentes.	La personne responsable de cette mission dispose de la légitimité et des ressources nécessaires pour coconstruire la gouvernance de l'IA.	La personne responsable de la gouvernance de l'IA rapporte directement au Comex avec le CTO (Chief Technical Officer).
MODÈLE DE GOUVERNANCE	Aucun modèle de gouvernance n'est choisi.	Recensement et comparaison des modèles de gouvernance actuels / ajout de thématiques sur l'IA digne de confiance à une gouvernance interne préexistante.	Création d'un comité de gouvernance interne afin de piloter la mise en place de la charte.	Le comité de gouvernance représente l'ensemble des métiers et zones géographiques de l'organisation, développe des protocoles de gestion des risques et des formations adaptées.	Le comité de gouvernance s'ouvre aux experts externes et communique sur ses décisions dans le rapport annuel. Il initie en parallèle une réflexion sur l'évolution du modèle de gouvernance.
PROTOCOLES ET OUTILS	Aucun outil de gouvernance de l'IA n'est identifié.	Recensement et comparaison des outils du marché.	Recensement et qualification des besoins des métiers ainsi que des risques liés à l'IA, notamment à travers une check-list.	Sélection et intégration des outils choisis dans une boîte à outils de gouvernance de l'IA et mise à la disposition des parties prenantes pertinentes.	Sensibilisation en interne et formation sur l'utilisation de la boîte à outils, elle-même mise à jour régulièrement.

— REMERCIEMENTS

« Félicitations à l'ensemble des contributeurs à ce guide très pratique, associé à la vision responsable et inclusive de l'intelligence artificielle que notre collectif Impact AI promeut. »

Agnes Van de Walle
— PRÉSIDENTE D'IMPACT AI
DIRECTEUR DIVISION PARTENAIRES ET STARTUPS
MICROSOFT FRANCE

Coordinateurs du cycle de travail sur la gouvernance d'une IA digne de confiance

Paul-Marie Carfantan	AI Ethics & Governance Consultant	DELOITTE FRANCE
Sarah El Marjani	Research & Partnership PMO	AXA

Membres du groupe de travail IA Responsable ayant participé aux différentes étapes de création du guide

Grégory Abisror	DELOITTE FRANCE
Jonathan Aigrain	AXA
Ismael Al-Amoudi	GRENOBLE EM
Teva Atani	DELOITTE FRANCE
Olivier Baes	MAIF
Daniel Bartolo	MAIF
Jean-David Benassouli	PWC FRANCE
Romain Bey	SUBSTRA
Gwendal Bihan	AXIONABLE
Emmanuel Bloch	THALES
Éric Boniface	SUBSTRA
Sandrine Bortone	CERCLE INTERELLES
Sylvie Caruso Cahn	SNCF RÉSEAU
Marcin Detyniecki	AXA
Hector Duport de Rivoire	MICROSOFT
Richard Eudes	DELOITTE FRANCE
Émilie Sirvent-Hien	ORANGE
Hélène Jeannin	ORANGE
Anne-Marie Jonquière	CEA
Claude Le Pape Gardeux	SCHNEIDER ELECTRIC
Eneric Lopez	MICROSOFT
Nicolas Marescaux	MACIF
Louis Montanié	PWC FRANCE
Thierry Notaire	MAIF
Bernard Ourghanlian	MICROSOFT
Florence Picard	INSTITUT DES ACTUAIRES
Yves Yota Tchoffo	DELOITTE FRANCE
Jean-Jacques Thomas	SNCF RÉSEAU
Camille Vaziaga	MICROSOFT

— LES MEMBRES D'IMPACT AI

ACCENTURE

ADECCO

ARTHUR D LITTLE

AXA

AXIONABLE

BOSTON CONSULTING GROUP

BLECKWEN

BOUYGUES

BURGUNDY SCHOOL OF BUSINESS

CAPGEMINI

CALLIOPE

CERCLE INTERELLES

CFE CGC

CGI

CRAFT AI

DATA EXPERT

DC BRAIN

DELOITTE

DEVOTEAM

DYNACENTRIX

ECONOCOM

EMERTON DATA

ENEDIS

EPITA

FAURECIA

FRANCE IS AI

GRANDE ÉCOLE DU NUMÉRIQUE

GRENOBLE ÉCOLE DE MANAGEMENT

GROUPE HERVÉ

GROUPE IGS

RATP GROUP

INSTITUT DES ACTUAIRES

INVIVO

LA POSTE

MACIF

MAGELLAN PARTNERS

MAIF

MAKE SENSE

MBA DMB

MICROSOFT

ORANGE

PWC

SCHNEIDER ELECTRIC

SICARA

SIMPLON.CO

SKEMA BUSINESS SCHOOL

SNCF

SOPRA STERIA

SUBSTRA FOUNDATION

TALAN

THALES

VEKIA

WAVESTONE

WILD CODE SCHOOL

— GLOSSAIRE

AGENT CONVERSATIONNEL (CHATBOT) [1]

Agent conversationnel qui dialogue avec son utilisateur (par exemple, les robots empathiques à disposition de malades, ou les services de conversation automatisés dans la relation au client).

ALGORITHME [1]

Description d'une suite finie et non ambigüe d'étapes ou d'instructions permettant d'obtenir un résultat à partir d'éléments ou données fournis en entrée.

APPRENTISSAGE MACHINE (OU APPRENTISSAGE AUTOMATIQUE, MACHINE LEARNING) [1]

Branche de l'intelligence artificielle, fondée sur des méthodes d'apprentissage et d'acquisition automatique de nouvelles connaissances par les ordinateurs, qui permet de les faire agir sans qu'ils aient à être explicitement programmés.

APPRENTISSAGE MACHINE SUPERVISÉ [1]

L'algorithme apprend à partir de données d'entrée qualifiées par l'humain et définit ainsi des règles à partir d'exemples qui sont autant de cas validés.

APPRENTISSAGE MACHINE NON SUPERVISÉ [1]

L'algorithme apprend à partir de données brutes et élabore sa propre classification qui est libre d'évoluer vers n'importe quel état final lorsqu'un motif ou un élément lui est présenté. Pratique qui nécessite que des instructeurs apprennent à la machine comment apprendre.

APPRENTISSAGE PROFOND (DEEP LEARNING) [2]

L'algorithme apprend en agençant et paramétrisant différentes couches de traitements non-linéaires inspirés par le fonctionnement du cerveau humain.

AUTONOMIE [2]

Capacité d'exécuter des tâches prévues à partir de l'état courant et des détections, sans intervention humaine selon la norme ISO 8373:2012.

BASE DE DONNÉES [2]

Recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement

1. CNIL, « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle », décembre 2017

2. Cigref et cabinet Alain Bensoussan Avocats, « Gouvernance de l'intelligence artificielle dans les entreprises », septembre 2016

accessibles par des moyens électroniques ou par tout autre moyen selon l'article L. 112-3 du Code de la propriété intellectuelle.

BIAIS ALGORITHMIQUE [3]

Déviation par rapport à un résultat attendu. Le résultat biaisé conduit souvent à des discriminations de personnes.

POUR ALLER PLUS LOIN : [4]
Les algorithmes sont conçus par des humains et à partir de données reflétant des pratiques humaines. Ce faisant, les biais peuvent être ainsi intégrés à toutes les étapes de l'élaboration et du déploiement des systèmes : dès l'intention qui préside à l'élaboration de l'algorithme en amont, pendant la réalisation du code informatique, celle du code exécutable, celle de l'exécution, celle du contexte d'exécution et celle de la maintenance.
Dans le cas d'un algorithme d'apprentissage, le biais sera souvent dû, non pas à l'algorithme proprement dit, mais à un déséquilibre (éventuellement représentatif des pratiques passées) des données utilisées lors de la phase d'apprentissage.

BIG DATA [1]

Désigne la conjonction entre, d'une part, la constitution d'immenses volumes de données devenus difficilement traitables par des systèmes de gestion classiques et, d'autre part, les nouvelles techniques permettant de traiter ces données, voire d'en tirer parti par le repérage de corrélations inattendues d'informations.

CHECK-LIST

Une check-list est un outil de référence, un document élaboré dans le but de ne pas oublier les étapes nécessaires d'une procédure pour qu'elle se déroule avec le maximum de sécurité.

COMITÉ ÉTHIQUE

Dans l'entreprise, le Comité éthique IA est l'instance de gouvernance responsable de la démarche éthique en IA. Il assure la définition du cadre de gouvernance et des processus de contrôle de conformité adaptés à l'entreprise. Il assure également la définition et le partage des objectifs, puis des plans d'action pour atteindre ces objectifs.

DÉVELOPPEMENT DURABLE ET BIEN-ÊTRE

Les systèmes d'intelligence artificielle devraient être utilisés pour soutenir des évolutions sociales positives et renforcer la durabilité et la responsabilité écologique.

1. CNIL, « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle », décembre 2017

3. Patrick Bertail, David Bounie, Stephan Clémençon, Patrick Waelbroeck, « Algorithmes : biais, , discrimination et équité », Télécom ParisTech, 2019

4. « Algorithmes : prévenir l'automatisation des discriminations », Défenseur des droits et CNIL, juin 2020

DIGNITÉ

Les systèmes d'intelligence artificielle devraient être les vecteurs de sociétés équitables en se mettant au service de l'humain et des droits fondamentaux, sans restreindre ou dévoyer l'autonomie des individus.

DONNÉE À CARACTÈRE PERSONNEL [2]

Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres selon l'article 2 de la loi Informatique et libertés.

EFFETS DE BORDS (SIDE EFFECTS)

Une fonction, considérée comme un nouvel environnement, modifie l'état d'une ou plusieurs variables et provoque ainsi des comportements non désirés.

ÉQUITÉ

Alors que l'on attend des résultats d'un algorithme de machine learning une neutralité parfaite, il arrive que ceux-ci soient biaisés. Un biais est un préjugé en faveur ou en défaveur d'une chose, d'une personne ou d'un groupe par rapport à un autre et qui est considéré comme injuste. L'IA équitable, débarrassée des biais, peut contribuer à réduire les inégalités et à bâtir une société plus inclusive.

EXPLICABILITÉ [5]

L'explicabilité consiste à comprendre ou donner à comprendre *comment* fonctionne un algorithme, *pourquoi* il prend telle décision et donne tel résultat, le but recherché, la garantie qu'il fonctionne correctement.

POUR ALLER PLUS LOIN :
L'explicabilité du fonctionnement ne garantit toutefois pas toujours celle des prédictions. La logique interne paraît parfois inexplicable et impénétrable : c'est l'effet « boîte noire » (black box). On considère aussi traditionnellement que certaines familles d'algorithmes sont plus susceptibles d'explicabilité que d'autres.
L'explicabilité se distingue de la transparence, l'interprétabilité et l'auditabilité :
– La transparence n'est qu'un moyen de donner à comprendre des décisions algorithmiques : elle traduit une possibilité d'accéder à une documentation, voire au code source des algorithmes et aux modèles qu'ils produisent. Plus largement, elle vise à fournir des informations

2. Cigref et cabinet Alain Bensoussan Avocats, « Gouvernance de l'intelligence artificielle dans les entreprises », septembre 2016

5. Définition partiellement inspirée du document de réflexion « Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier », Laurent Dupont, Olivier Fliche, Su Yang, ACPR, juin 2020

concernant : a) le fait qu'une organisation utilise un système d'IA dans un processus de prise de décision ; b) les objectifs prévus du système d'IA et la façon dont le système d'IA sera et peut être utilisé ; (c) les types d'ensembles de données utilisés par le système d'IA ; et (d) des informations utiles sur la logique visée.
- L'auditabilité caractérise la faisabilité pratique d'une évaluation analytique et empirique de l'algorithme, et vise plus largement à obtenir non seulement des explications sur ses prédictions, mais aussi à l'évaluer selon les autres critères indiqués précédemment (performance, stabilité, traitement des données).
- La distinction entre explicabilité et interprétabilité est âprement débattue : le terme d'explicabilité est souvent associé à une compréhension technique et objective du fonctionnement d'un algorithme (et serait donc adapté à la perspective d'une mission d'audit), tandis que l'interprétabilité semble davantage liée à un discours de nature moins technique (et viserait donc avant tout le consommateur ou l'individu impacté par l'algorithme).

GESTION DES VERSIONS (VERSIONING)

La gestion des versions d'un document, d'un algorithme, d'un modèle appris par un algorithme, etc., consiste à mettre en œuvre des moyens permettant de mémoriser et de retrouver ou reconstruire les différentes versions de l'élément considéré.

POUR ALLER PLUS LOIN :
Selon les besoins d'une application d'intelligence artificielle, il peut être nécessaire de gérer non seulement les différentes versions des algorithmes, mais aussi les versions successives des modèles appris et des données utilisées lors de l'apprentissage et l'exécution des modèles. Dans la pratique, un compromis entre la capacité de restauration de nombreuses versions et les ressources de stockage nécessaires doit être trouvé. Il est également nécessaire de tenir compte des règles relatives à la conservation des données privées ou confidentielles. Les principes de gouvernance des données et de transparence/traçabilité peuvent donc parfois s'opposer.

GLISSEMENT DE CONCEPT OU DE MODÈLE (CONCEPT DRIFT / MODEL DRIFT)

Le glissement de concept (ou de modèle) désigne la situation dans laquelle un modèle appris dans un contexte donné est utilisé dans un contexte « décalé », mal représenté par le modèle.

POUR ALLER PLUS LOIN :
Dans la plupart des cas, un tel glissement se produit lorsque le modèle n'incorpore pas toutes les données susceptibles d'influencer le résultat, soit parce que ces données sont non-mesurables

ou non-mesurées, soit parce qu'ayant été très stables par le passé, leur influence n'a pas pu être constatée. Dans tous les cas, la fiabilité du résultat peut en être significativement altérée.
Par exemple, un modèle de prédiction du prix d'un produit de consommation peut avoir été établi dans un contexte concurrentiel tel que seuls la marque, le modèle et la date initiale de mise sur le marché du produit se sont avérés pertinents. Si le contexte concurrentiel change, les prix et conditions pratiqués par des concurrents peuvent modifier la donne. Les résultats fournis par le modèle initial risquent de se révéler inexacts.

GLISSEMENT DE DONNÉES (DATA DRIFT)

Le glissement de données désigne la situation dans laquelle la distribution des données fournies à un algorithme évolue au cours du temps.

POUR ALLER PLUS LOIN :
Dans le cas d'un algorithme qui applique un modèle résultant d'un apprentissage, le modèle appris n'est donc pas représentatif des nouvelles données et la qualité de ses prédictions est susceptible de se détériorer.
Par exemple, une solution de machine learning pour réguler le chauffage d'un appartement peut avoir été entraînée avec une distribution « habituelle » des données de températures extérieures.
Si des températures extrêmes se présentent, le réglage du chauffage risque d'être sous-optimal.

GOUVERNANCE [6]

La traduction et l'inscription de principes dans une réalité concrète. La gouvernance se manifeste par un ensemble de processus adoptés par une organisation, réglementations, lois et institutions qui influent sur les orientations prises par l'organisation dans différents domaines (administration, contrôle, politique commerciale, marketing…). Elle résulte d'une dynamique mêlant acteurs internes (salariés…) et externes (clients, associations de consommateurs…), agents économiques ou non, capables d'exercer différents types de pression et considérés comme des parties prenantes de l'organisation à part entière.

POUR ALLER PLUS LOIN :
La gouvernance réfère à l'implication de ces parties concernées aux actions prises par une organisation socialement intégrée (c'est-à-dire, non considérée comme un univers clos mais prise dans un contexte de relations étroites et d'interdépendances) afin d'assurer la maîtrise de tous les aspects lui permettant d'atteindre ses objectifs. Pour une entreprise, ceux-ci restent naturellement d'assurer sa viabilité, son développement et son efficacité

6. Définition partiellement inspirée de Sophie Boutillier, Castilla Ramos Beatriz, « Gouvernance et responsabilité sociale des entreprises internationales : l'exemple d'une entreprise américaine implantée au Mexique », Marché et organisations, 2009/2 N° 9, p. 89-118

économique, mais englobent également la prise en considération de nombreux autres paramètres. La gouvernance d'une organisation peut donc être orientée vers un ensemble complexe d'objectifs intégrant les principes de l'IA responsable sans que la question de la réalisation du profit ou de sa pérennité ne soit remise en cause. Plus concrètement la gouvernance peut prendre en compte l'intégration de l'IA dans les processus (métiers, décisionnels) de l'organisation et l'impact de cette intégration, la gestion des risques associés à l'IA et les procédures opérationnelles associées, la présentation des normes éthiques et professionnelles promues (ainsi que les mesures d'application ou de recours en cas de manquement). Elle peut s'accompagner de l'adoption volontaire d'un ensemble d'exigences adaptées à l'activité de l'organisation, formalisées et documentées (niveau minimal d'explicabilité, de transparence, de traçabilité…) et permettant de répondre à la diversité des besoins et pressions. Ainsi les salariés sont également des citoyens et des consommateurs, l'IA véhicule un certain nombre de craintes capables de canaliser les énergies et d'alimenter des résistances (peur de perdre son travail, d'être surveillé…).

GOUVERNANCE DES DONNÉES

Ensemble des mesures qui visent à s'assurer de la qualité et de l'intégrité des données utilisées ainsi que de leur pertinence dans le domaine d'application considéré. Il s'agit aussi d'établir des protocoles d'accès à ces données et des règles pour les traiter qui garantissent le respect de la vie privée.

HYPERPARAMÈTRES

Les hyperparamètres désignent des paramètres dont les valeurs, fournies à un algorithme d'apprentissage, influencent le processus d'apprentissage.

POUR ALLER PLUS LOIN :
Ces paramètres dépendent de l'algorithme utilisé. Ils peuvent porter sur le type et la taille maximale du modèle appris ou sur l'agencement et le dimensionnement des étapes successives du processus d'apprentissage. En pratique, le choix de valeurs pour ces paramètres a une influence sur la qualité des modèles et sur les ressources nécessaires à l'apprentissage et à l'utilisation des modèles.

INTELLIGENCE ARTIFICIELLE (IA)

L'intelligence artificielle (IA) est un domaine vaste et pluridisciplinaire dont l'unité tient dans l'ambition initiale de faire reproduire par les machines des compétences généralement attribuées à l'être humain, telles que le raisonnement ou l'apprentissage.

POUR ALLER PLUS LOIN :
L'IA comprend d'une part des technologies cœur (composants) et d'autre part des applications ou systèmes intelligents, résultant de l'intégration d'un ou plusieurs composants technologiques d'IA et d'autres composants techniques non IA spécifiques au domaine métier concerné. L'IA consiste avant tout en des applications concrètes : reconnaissance faciale, traitement automatisé du langage, vision par ordinateur, etc., dont le dénominateur commun est la finalité, par exemple la possibilité d'effectuer des tâches ou un ensemble de tâches complexes beaucoup plus rapidement qu'un être humain grâce à des capacités de traitement et de calcul supérieures. De nombreuses applications d'intelligence artificielle reposent sur la notion d'apprentissage. Afin de permettre à un système d'apprendre, il faut mettre en commun des bases de données. Le *Big data* est donc un enjeu non négligeable. Néanmoins, il existe plusieurs types d'apprentissage (supervisé, non supervisé, cf. Apprentissage).

IA DIGNE DE CONFIANCE [7]

La confiance est considérée comme un prérequis indispensable à l'acceptation et à l'adoption de l'intelligence artificielle, une confiance qui non seulement s'acquiert par la pédagogie mais qui s'entretient aussi dans la durée.

POUR ALLER PLUS LOIN :
Une IA digne de confiance présente les trois caractéristiques suivantes, qui devraient être respectées tout au long du cycle de vie du système : a) elle doit être licite, en assurant le respect des législations et réglementations applicables ; b) elle doit être éthique, en assurant l'adhésion à des principes et valeurs éthiques ; c) elle doit être robuste, sur le plan tant technique que social car, même avec de bonnes intentions, des systèmes d'IA mal conçus peuvent causer des préjudices involontaires.

IA RESPONSABLE [8]

L'obligation aux acteurs qui proposent des solutions basées sur cette technologie de tenir compte de leurs impacts économiques, sociétaux et environnementaux, dès le stade de la conception.

POUR ALLER PLUS LOIN :
La notion d'IA responsable reflète les préoccupations relatives aux nouveaux enjeux éthiques qui se posent en raison des avancées, du déploiement et de la diffusion de l'IA dans de nombreux secteurs d'activité et dans le quotidien. Elle témoigne de la prise de conscience de nombreux acteurs (communauté scientifique, organisations…) du potentiel de l'IA. Elle marque la volonté de traiter et d'anticiper les risques et dérives potentielles par la mise en place de garde-fous, afin de s'assurer que l'IA soit mise au service de l'humain.

7. Définition partiellement inspirée du document de réflexion du groupe d'experts indépendants de haut niveau sur l'intelligence artificielle constitué par la Commission Européenne « Lignes directrices en matière d'éthique pour une IA digne de confiance », juin 2018

8. Définition partiellement inspirée de Duong Quynh Lien, « La responsabilité sociale de l'entreprise, pourquoi et comment ça se parle ? », *Communication et organisation*, N°26, 2005

Elle s'accompagne d'une mission de sensibilisation et d'information pertinente qui vise à lever les freins qui retarderaient l'adoption de services et produits d'IA. Les organisations qui développent, déploient ou utilisent des systèmes d'IA doivent surveiller la mise en œuvre de ces systèmes et agir pour atténuer les conséquences (intentionnelles ou non) de ces systèmes d'IA qui seraient incompatibles avec les objectifs éthiques de bienfaisance et de non-malfaisance, tout comme avec les autres principes de l'énoncé de politique pour une IA responsable. On constate à travers notre recherche que l'IA responsable entretient une filiation étroite avec la RSE (Responsabilité sociétale de l'entreprise). En effet, elle témoigne du même souci de replacer l'entreprise dans son environnement, de répondre de l'effet de ses actions sur celui-ci et d'entrer en adéquation avec lui au regard des attentes vis-à-vis de son comportement ou de ses activités. Cet environnement social avec lequel elle est en interaction permanente est très diversifié. Il englobe différentes parties prenantes marquées par des rapports de force et de concurrence. L'IA responsable s'inscrit donc au carrefour de sensibilités et de courants de pensée mêlant responsabilité morale, recherche de facteur concurrentiel ou différenciant, souci de la réputation et de l'image, nécessité d'intégrer la nouveauté et l'innovation dans une démarche gestionnaire, organisationnelle et de communication. L'ensemble s'inscrit dans le cadre de normes et de règles que l'organisation contribue d'ailleurs elle-même parfois à construire, à diffuser et à institutionnaliser. Pour faire valoir sa démarche et être en capacité d'influence elle recourt à différents moyens : publicité, chartes ou codes de conduite, rapports ou livres blancs, conférences ou ateliers, participation à des instances de labellisation ou de normalisation, recours à la certification…

MACHINE LEARNING [2]

Apprentissage automatique des machines, branche de l'intelligence artificielle permettant aux machines d'effectuer des tâches pour lesquelles elles ne sont pas explicitement programmées, en apprenant des modèles à partir de données.

OPEN SOURCE

Une application en *open source* est un programme informatique dont le code source est distribué sous une licence permettant à quiconque (sous des conditions définies par la licence) de lire, modifier ou redistribuer ce logiciel.

RESPONSABILITÉ

Le développement de l'intelligence artificielle doit s'accompagner de garanties suffisantes pour minimiser le risque de dommages qu'elle peut causer. Il convient de mettre en place des mécanismes à cette fin et de les soumettre à l'obligation de rendre des comptes.

2. Cigref et cabinet Alain Bensoussan Avocats, « Gouvernance de l'intelligence artificielle dans les entreprises », septembre 2016

ROBOT [2]

Un robot présente les caractéristiques suivantes :
– une machine intelligente dotée d'un module d'intelligence artificielle ;
– une capacité à prendre des décisions en ne se réduisant pas à obéir à des automatismes ;
– des capacités d'apprentissage ;
– une situation de mobilité dans des environnements privés et publics ;
– une capacité à agir de manière coordonnée avec des êtres humains.

ROBOT INTELLIGENT [2]

Robot capable d'exécuter des tâches par détection de son environnement et/ou par interaction avec des sources extérieures et adaptation de son comportement selon la norme ISO 8373:2012.

ROBUSTESSE

Une IA digne de confiance nécessite des algorithmes suffisamment sûrs et fiables pour gérer les erreurs ou les incohérences à toutes les étapes du cycle de vie d'un outil d'intelligence artificielle.

TRANSPARENCE

Les algorithmes d'apprentissage sont dits transparents lorsque le fonctionnement interne des modèles d'apprentissage est rendu public ; ils répondent à l'exigence d'explicabilité quand ces mêmes éléments restent intelligibles, c'est-à-dire qu'il est possible de comprendre à quelles règles ils obéissent.

VÉRITÉ DE TERRAIN (GROUND TRUTH)

La vérité de terrain désigne les données acquises par la mesure directe sur le terrain, par opposition aux données calculées par un algorithme.

POUR ALLER PLUS LOIN :
Les données de terrain passées constituent une entrée essentielle d'un algorithme d'apprentissage supervisé, permettant d'établir un modèle d'estimation ou de prédiction de certaines de ces données. Dans la pratique, il convient d'évaluer la qualité du modèle (dès sa construction et idéalement – ce qui n'est pas toujours possible – tout au long de son cycle de vie) en comparant les prédictions réalisées à des données de terrain dont ne disposait pas l'algorithme d'apprentissage.

2. Cigref et cabinet Alain Bensoussan Avocats, « Gouvernance de l'intelligence artificielle dans les entreprises », septembre 2016

— TABLE DES FIGURES

— BIBLIOGRAPHIE

ARBORUS ORANGE. « Charte internationale pour une I.A. inclusive », avril 2020.

ASCHIÉRI, GÉRARD. « Sciences et société : les conditions du dialogue », 2020.

ATAWAO CONSULTING, CGET, TECH'IN, DGE, ET PIPAME. « Intelligence artificielle : état de l'art et perspectives pour la France : rapport final », 2019.

AXA. « Understanding and Mitigating Bias », 2019.

AXA. « Regulating Machine Learning: where do we stand? », 2019.

BECK, ULRICH. *La société du risque : sur la voie d'une autre modernité.* Flammarion, 2008.

BERTAIL, PATRICE, DAVID BOUNIE, STEFAN CLÉMENÇON, ET PATRICK WAELBROECK. « Algorithmes : biais, discrimination et équité », 2019.

BOSTON CONSULTING GROUP, ET MALAKOFF MÉDÉRIC HUMANIS. « IA et capital humain : l'humain est l'avenir de l'IA, deuxième édition », 2019.

BOUTILLIER, SOPHIE, ET BEATRIZ CASTILLA RAMOS. « Gouvernance et responsabilité sociale des entreprises internationales : l'exemple d'une entreprise américaine implantée au Mexique ». Marche et organisations N° 9, n° 2 (2009) : 89 118.

CERCLE INTERELLES. « Cercle InterElles, Actes complets du colloque 2019 », 2019.

CIGREF, ET CABINET ALAIN BENSOUSSAN AVOCATS. « Livre Blanc CIGREF « Gouvernance de l'Intelligence Artificielle dans les entreprises » », septembre 2016.

CNIL, ET DÉFENSEUR DES DROITS. « Algorithmes : prévenir l'automatisation des discriminations », 2020.

CNIL, ET ÉTHIQUE ET NUMÉRIQUE. « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle », 2017.

DELOITTE. « Thriving in the era of pervasive AI, Deloitte's State of AI in the Enterprise, 3rd Edition », 2020.

« Talent and workforce effects in the age of AI, Insights from Deloitte's State of AI in the Enterprise, 2nd Edition », 2020.

DUPONT, LAURENT, OLIVIER FLICHE, SU YANG, ET ACPR. « Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier », 2020.

SHAPING EUROPE'S DIGITAL FUTURE - EUROPEAN COMMISSION. « Ethics Guidelines for Trustworthy AI ». avril 2019.

EUROPEAN UNION. « Digital Single Market - Artificial Intelligence ». Shaping Europe's digital future - European Commission, mars 2018.

FRANCE INTELLIGENCE ARTIFICIELLE. « La stratégie IA en France ». Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, mars 2017.

GOODFELLOW, IAN J., JONATHON SHLENS, ET CHRISTIAN SZEGEDY. « Explaining and Harnessing Adversarial Examples ». *arXiv:1412.6572 [cs, stat]*, mars 2015.

GROUPE D'EXPERTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE., ET EUROPEAN COMMISSION. DIRECTORATE GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY. *Lignes directrices en matière d'éthique pour une IA digne de confiance.* Publications Office, 2019.

GUCHET, XAVIER. « L'éthique des techniques, entre réflexivité et instrumentalisation ». Revue francaise d'ethique appliquee N° 2 (juin 2016) : 8 10.

« High-Level Expert Group on Artificial Intelligence | Shaping Europe's digital future ».

IEEE. « Ethically Aligned Design (EAD1e): A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems », 2019.

IMPACT AI. « Un engagement collectif pour un usage responsable de l'intelligence artificielle », 2019.

INRIA. « Inria présente les projets de sa mission covid », inria.fr, mai 2020.

INSTITUT MONTAIGNE. « Algorithmes : contrôle des biais S.V.P. » Institut Montaigne, mars 2020.

« Instruments juridiques de l'OCDE : recommandations du Conseil sur l'intelligence artificielle », 22 mai 2019.

JOBIN, ANNA, MARCELLO IENCA, ET EFFY VAYENA. « The Global Landscape of AI Ethics Guidelines ». Nature Machine Intelligence 1, n° 9 (septembre 2019) : 389 99.

« L'efficacité d'un logiciel censé prédire la récidive à nouveau critiquée ». Le Monde.fr, janvier 2018.

« Manuel de droit européen en matière de protection des données », 2018.

LAROUSSERIE, DAVID. « Coronavirus : comment l'intelligence artificielle est utilisée contre le Covid-19 ». Le Monde.fr, mai 2020.

ORANGE. « Orange dévoile sa raison d'être co-construite », 10 décembre 2019.

PEDRO A. ORTEGA, VISHAL MAINI, AND THE DEEPMIND SAFETY TEAM. « Building Safe Artificial Intelligence: Specification, Robustness, and Assurance ». Medium, septembre 2018.

QUYNH LIEN, DUONG. « La responsabilité sociale de l'entreprise, pourquoi et comment ça se parle ? » Communication et organisation, n° 26 (janvier 2005): 26 43.

RUF, BORIS, CHAOUKI BOUTHAROUITE, ET MARTIN DETYNIECKI. « Getting Fairness Right: Towards a Toolbox for Practitioners », mars 2020.

SERMONDADAZ, SARAH. « Intelligence artificielle : la reconnaissance faciale est-elle misogyne et raciste ? » Sciences et Avenir, 6 mars 2018.

SHAPING EUROPE'S DIGITAL FUTURE - EUROPEAN COMMISSION. « Ethics Guidelines for Trustworthy AI », juillet 2020.

SHAPING EUROPE'S DIGITAL FUTURE - EUROPEAN COMMISSION. « The Digital Economy and Society Index (DESI) », février 2015.

STEIN, SHIRA, JACOBS JENNIFER. « Cyber attacks hits U.S Health Agency Amid Covid-19 outbreak », Bloomberg.com, mars 2020.

VILLANI, CÉDRIC, MARC SCHONENAUER, YANN BONNET, CHARLY BERTHET, ANNE-CHARLOTTE CORNUT, FRANÇOIS LEVIN, ET BERTRAND RONDEPIERRE. « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », 2018.

WORLD ECONOMIC FORM AND DELOITTE. « Navigating Uncharted Waters : A Roadmap to Responsible Innovation with AI in Financial Services », 2020.

CONCEPTION ET COORDINATION : Paul-Marie Carfantan, Sarah El Marjani

COMITÉ DE RÉDACTION : Grégory Abisror, Olivier Baes, Daniel Bartolo, Gwendal Bihan, Emmanuel Bloch, Eric Boniface, Sandrine Bortone, Paul-Marie Carfantan, Sylvie Caruso Cahn, Hector Duport de Rivoire, Marcin Detyniecki, Sarah El Marjani, Emilie Hien, Hélène Jeannin, Anne-Marie Jonquière, Claude Le Pape Gardeux, Louis Montanié, Thierry Notaire, Roxana Rugina, Camille Vaziaga, Yves Yota Tchoffo

REWRITING : Isabelle Raoul-Chassaing

CRÉDITS PHOTOS : p. 25 © hobby/Unsplash; p. 35 © Vlad Hilitanu/Unsplash; p. 43 © Amanda Dalbjorn/Unsplash; p. 47 © Charles Deluvio/Unsplash, © Theodor Lundqvist/Unsplash; p. 51 © Andreea Popa/Unsplash; p. 55 © Mahdis Mousavi/Unsplash; p. 67 © Marius Masalar/Unsplash; p. 73 © Victor Garcia/Unsplash; p. 79 © Alexander Sinn/Unsplash; p. 83 © Max Langelott/Unsplash

DESIGN & PRODUCTION : Aurélien Saublet

IMPRESSION : Imprigraphic

IMPACT AI www.impact-ai.fr - décembre 2020


IMPACT AI